Journal of Theoretical and Applied Information Technology
15th July 2019. Vol.97. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CONDITION-BASED MAINTENANCE USING DATA MINING TECHNIQUES ON INTERNET OF THINGS GENERATED DATA

[1]FRANKY RADIANSYAH, [2]TUGA MAURITSIUS
[1,2]Information System Management Department,
BINUS Graduated Program – Master Information System Management
Bina Nusantara University
Jakarta 11480 Indonesia

E-mail: [1]franky.radiansyah@binus.ac.id; [2]tuga.mauritsius@binus.ac.id;

ABSTRACT

Heavy Equipment Industry have various business counterparts, including mining industries, infrastructure contractors, and as well as any kind of manufactures. Currently companies in the similar business are working hard on how to optimize the maintenance activities on their heavy equipment. Maintenance of those equipment could be very crucial to the business continuity. This paper provides an alternative to optimize such an activity through an approach called condition-based maintenance. We conducted our research in one international heavy equipment rental company based in Singapore and has a branch in Indonesia. The company's core business is on heavy equipment rental including Excavator. The research focused on utilizing data generated by sensors attached to the Excavator with the main aim is to predict the Remaining Useful Life (RUL) of Oil Grease Pump which is a crucial component of the Excavator. We used some machine learning techniques such as Linear Regression, Decision Tree Regression, and Random Forest methodology to build models to predict the RUL. The results from each models were compared each other to gain a deeper insight on the predictive ability of each model using the data provided. It turns out that the linear regression model gives the highest predictive accuracy with 61% of RMSE.

Keywords: *Machine Learning, Condition Based Maintenance, Predictive Maintenance*

1. INTRODUCTION

Indonesia is a rich country with a lot of natural resources spreading on many island, from Sabang to Merauke. In order to maximize the use of its natural resources, infrastructure development is a strategic plan made by the government of Republic of Indonesia during 2014 - 2019 periode. Because of that, the heavy equipment industry which is part of Various Industry Sector of Indonesian Stock Exchange, becoming the ace industry to determine the succeed for every infrastructure development in Indonesia.

Heavy Equipment Industry have various of business area, such as mining or infrastructure industry, or a service provider for transportation and heavy equipment to the clients who need. The objective of this research is international heavy equipment company which also have Indonesia branch, where their focus are to rent those heavy equipment and transportation services to the government and public including either the maintenance services for those rented heavy equipment (managed service).

One of the main focus in every services industry is Service Lever Agreement (SLA). This is a base of what needs to follow and to be done related with service performance, additionally to increase the level of trust between the service provider and client. One of the SLA point that have to be known is presentation of the heavy equipment up-time have to follow the standard.

There are known 3 kind of maintenance, Reactive or Corrective Maintenance, Preventive Maintenance (PvM), and Predictive maintenance (PdM). These days many company choose to do reactive and preventive maintenance, where they only do maintenance when the damage is happening (reactive) else to prevent the damage (preventive). Actually both are not efficient or effective because reactive maintenance is waiting for the damage to be happen which impacting the SLA, and preventive maintenance is to preventing

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the damage even though it may not be required and may impacting the unnecessary cost.

According to the company data on this research, we found that there are some cases of spare parts replacement happening before the schedule, and impacting to non measureable expense in other words unexpected expense, causing the company spending cost earlier than expected schedule. We took an example where an injector spare part expiry up-time is 22000 hours, but the actual up-time was 9000, from the known gap clearly it will impact to downtime, where machine will stop operating until the spare part got replace.

As a method to achieve a better level of effectiveness and efficiency , therefore required a predictive maintenance system or condition based maintenance. Through these system, company will have a better understanding of what will happen and need to optimize the maintenance expense plan and consistently maintain the SLA.

One of the requirement to use the condition based maintenance is quality of the data, which represent the situation and condition of the heavy equipment where this will be implement. Thing which will be utilized to improve transparency of the real situation and condition of those heavy equipment is to implement the Internet of Things (IoT) tools which give a hope to be the source of the data in order to make a decision.

IoT is used as a censors or alarm who collecting data for situation and condition of excavator functionality such as temperature, rotation, lift capacity, current weight, on progress job and etc. Those data spread on every excavator and automatically upload to the cloud drive. Even though the storage is using the cloud based, it does have a limitation, where the more data growing the more cost will be spend for it. In fact those data stored and passive at the storage and only use when a report need to generate. Knowing these situation is quite unfortunate where the data is not turn into something more use full. Therefore this research is interested to process those data to knowing if those data could be use to condition based maintenance prediction.

According to the data, this research focus will be one of the most important spare part of sub-system excavator which is lubrication system. Part of the lubrication system on the excavator is the oil grease pump, where its spare-part impacting the performance of the excavator. Therefore the historical data of this spare-part will be used as one of research composition. And to be known that there are many censors on the excavator, so there will be many variables involved during the data processing on this research.

After the data is available, data mining has role during data processing of condition based maintenance implementation. And for predictive maintenance, there are some suitable data mining technique which can be use, such as: Linear Regression, Random Forest Regression and Decision Tree Regression

According to the background, therefore writer conclude the problem formulation based on the problem identification:

1. Is it possible to use IoT-based historical data to predict the RUL of oil grease pump using some data mining techniques such as : Liner regression, decision tree regression and random forest.
2. How to manage the high dimensions data in the RUL Prediction system development for oil pump grease attached on excavator?
3. How is the accuracy of those data mining techniques in predicting the RUL of the oil pump grease?
4. How to utilize the outcome of the RUL prediction into Decision Support System development of Condition Based Maintenance?

Through the explained problem formulation, therefore the objective of this research is to investigate and explore the use of IoT data for condition Based Maintenance purpose.

1. Perform a efficiency testing of machine learning technique for RUL predictive against the existing IoT data.
2. Perform a dimensional reduction against the existing data set to increase efficiency and algorithm performance.
3. Perform a comparison between the prediction outcome from each of machine learning technique against the RUL from the oil pump grease of the excavator.
4. Design a Decision support system based on the outcome from RUL prediction for condition based maintenance purpose.

2. LITERATURE REVIEW

2.1 PREDICTIVE MAINTENANCE

Predictive Maintenance have some requirement which is historical data availability. To enrich identification and detection of the failure, therefore required many data which can be associated to a problem or failure such as parameters and machine condition. In order to support data collection, the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

censor tools of IoT is used as the data source to do a prediction combined with Big Data platform.

2.2 INTERNET OF THINGS

The words of Internet of Things itself becoming a term for every tolls which related with internet. This could be machine, smart tools, indeed a building and factory (Georgakopoulos & Jayaraman, 2016). However, does it have a benefits if it just connected with the internet? Technology evolution is now used IoT as the real time censor where every condition can be captured by the IoT tools and send through the internet then store it into a storage either structured or non structured, with a traditional system such as SQL or NOSQL.

In this case, every machine and heavy equipment has various of censors to capture real conditions from a machine and those heavy equipment then stored into a system, after that those captured data will be processed to produce a prediction. There are many censors and various type of data which captured with a different level of data granularity. Therefore, required a method to do data processing before going forward it has a further data processing.

2.3 DATA MINING AND BIG DATA

Data Mining is a term used to outline an invention of knowledge inside a database. Data Mining is a process with a statistic technique, math, artificial intelligence, and machine learning to extract and identify information to get its benefits and correlated knowledge from various big database. (Turban, dkk. 2005).

According to Kusrini dan Emha Taufiq, 2009, there are important things need to understand and followed for data mining such as;

1. Data Mining is automation process against the existing data.
2. Data which are going to be processed is huge.
3. The purpose of data mining is to get a schema which could give a good indication.

Therefore, can be conclude that Data Mining is analysis process against data to extract a hidden information on some big data which stored when running the company business.

2.4 MACHINE LEARNING

Machine Learning is technique of Information Technology Infrastructure management which allows to create a decision making process with a system model based from the existing data on its system, so that the produced model can be updated, or creating a new model based on the specific requirement. In other words, that the used of Machine Learning for Information Technology Infrastructure management , we could dg a knowledge directly from the real time behaviour of a system (Berral et al., 2012).

Machine Learning start with processing the existing data, with a purpose to get interconnected information and to decide which can be the attribute, then creating a model that can be used to explained the actual condition of a system and led to a decision based on that model.

Generally how this Machine Learning work is processing a collection of data so called data set from a system by determining values from the system, determining the attribute and respond, then creating a model based from the mentioned values, so that when there are new upcoming data, the expected value will aligned with expectation of a produced model.

According to *Josep LL. Berral* and colleague *(Berral et al., 2012)*, implementation technique of *Machine Learning* divided by some approach such as :

- *supervised learning*, like: *classification* or *regression*
- *unsupervised learning* (to get understanding on a relation between the input data), such as ti: *clustering* (to get understanding on similarities of data input) or *reinforcement learning* (to select the best decision based on the previous event)

The purpose of Supervised Learning is to predict an outcome based on the input; while on the outcome of unsupervised learning is no outcome measure, but its purpose is to explain associate and schema between the input data. Although both approach is plausible to use input data to create a prediction as an output *(Hastie, Tibshirani, & Friedman, 2009).*

2.5 ALGORITMA

2.5.1 Linear Regression

Linier Regression is the best and simple algorithm to process a numeric data to produced a prediction, and widely used on many statistic calculation for past decades, although have some lack of its liner. Therefore, for data with a non-linier model, the best approach to get a straight line is by using the least-mean-squared difference. It's true that the straight line is not always perfect, but a model with linier approach is a first stage to move into further stage with a complex machine learning method, (Torres Viñals, 2010)).

Linier equation can be written like,

$$y = w_0 + w_{1a1} + w_{2a2} + \ldots + w_{kak}$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Where,
$y = Output\ a1, a2, a3, \dots, ak = Atribut/feature$
$w0, w1, w2, \dots, wk = koefisien$

According to Hasan (2008), multiple linier analysis is connecting relative variables (Y) by one or more open variables but still pointing to linier aligned diagram. By adding open variables hopefully could explain more about existing connection characteristics, although there are neglected variable. Common similarities multiple linear regression equation can be explained as :

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k + e$$

Noted :
Y = bound variable
b1, b2, b3, ..., bk = coefficient regression
X1, X2, X3, ..., Xk = open variable
e = disturbance terma,
meaning the values of the other variable which not part of the equation. This value usually ignored from calculation.

2.5.2 Decision Tree Regression

Decision Tree Regression is one of method which using a decision tree that produced by recursive insulation algorithm. This method analyzing a group of data by separate them into some of child group gradually.

Decision Tree algorithm and Decision Tree Regression has proposed by many writer. Most of the used algorithm is ID3 (Quinlan 1986) which developed into new algorithm C4.5 and C5.

Decision Tree Regression also explained how the relationship between respond variable and open variables. The difference is that on Decision Tree Regression method, the effect of open variables and its response performed by a group of observation based on the open variables, so that the interpretation outcome of this method can be easy performed. This is because of the open identification to the open variables was performed in every data child group, not in a single data set like a normal regression. Besides, Decision Tree Regression able to overcome an outlier. A simple statistic calculation is used on Decision Tree Regression so that become an advantage of this method.

2.5.3 Random Forest Regression

The first Random Forest model was introduced and start becoming a discussion by Breiman article (2001) aired on Machine Learning journal to build a predictor with a group of decision tree which randomly developed on a child data group.

Random Forest model is advance model of Decision Tree. If Decision Tree can be used either for regression problems or classification, Random Forest also can be used for the same purposed. When it used for classification problem, random forest produced a class vote for every tree, which are grouped based on majority vote. When used for a regression problem, the prediction outcome from every tree on X target performed by doing its average, like equation example below (Hastie et al., 2009):

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x; \Theta_b)$$

2.6 MODEL EVALUATION

In order to create a prediction, try to minimalize the uncertainty. In the other words, the prediction intend to minimalize a forecast error. Forecast error can be measured with Mean Absolute Error (MAE), MAE is score average of Absolute Error from an error forecast and MAE is the average from a squared forecast error (Subagyo, 1986), with below formula :

- Mean Absolute Error (MAE)
  $MAE = \frac{\Sigma \mid Yt - \hat{Y}t \mid}{n}$
- Mean Squared Error (MSE)
  $MSE = \frac{\Sigma \mid Yt - \hat{Y}t \mid 2}{n}$

Where :
Yt : actual data
$\hat{Y}t$: forecast data calculated from the used time (t) on a model
n : number of the output of forecast data

The real error will not be averaged as the big or small measurement of an error, with various positive and negative score, so if it's sum up, the error score will become small, impacting to a deviation from the real forecast which is actually big but seems small because of the sum up of the error, causing the positive error removed by a huge negative error. To avoid this condition, error have to be an absolute number or squared then do the average. (Subagyo, 1986).

Root Mean Squared Error or RMSE as an estimator is one of various ways to measure amount with different estimator from the real score of the predicted score. As a loss function, RMSE also called as a squared loss error. MSE measure the average of error from four square. Error is an amount that estimated from projected quantity. The difference happen because of the random score or different estimator cannot produce a score to be

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

used as an information which could create a different estimation with a better level of accuracy. Mean Square Error used to calculate an error level from two outcome of trial model, or if its implemented on a data mining, to measure the error level from the calculation of analysis which is using a specific method between training and testing data, the RMSE formula will be like :

$$RMSE = \sqrt{\frac{1}{m \times n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} [f(i,j) - g(i,j)]^2}$$

2.7 DECISION SUPPORT SYSTEM

Decision Support System developed since 1970. Decision Support System (DSS) supported with information system which can help people to improve their performance to make a decision. In other words Decision Support System (DSS) is part of information system with a computer based to support the business or decision making activities in a company or organization.

According to article from Haettenschwiler, he's telling that there are three type of DSS :

- Passive DSS, is a system to support a decision making process, but can't give either decision advice or solution,
- Active DSS, can give a decision advice or a clear solution,
- Cooperative DSS, is possible to repeatedly process between human and system against the achievement of consolidate solution. Decision maker can modify, completing or to fix the available decision from the system to do a validation.

The three basic component of DSS design is :

- Database or knowledge base,
- A model covering a decision context and user criteria,
- User interface (UI).

In addition to these three basic components, the users themselves also include important components of the design or design of the DSS.

There are 4 stages that need to be considered in the Decision Support System (DSS), including:

- Intelligence : Discover a condition that will be used to call a decision.
- Design : Develop and analysis a possible an alternative action from a solution.
- Choice : Select an action between the two previous stage.
- Implementation : Using the selected that has been chosen on a decision-making situation.

2.8 PREVIOUS STUDIES

Data Mining and Machine Learning for condition-based maintenance based According, Manzeni, Pascarella, Patella, and Sassi (2017) trying to design a new framework for system condition based maintenance that hopefully can perfecting two common maintenance method which is corrective and preventive. Here, writer trying to make a framework which covering various method to get a good prediction outcome of clustering, association rule that can see a connection schema of a message problem and failure, also classification from every problem message. On the design, writer also trying to transparently see a model, because an accurate model without transparency and hardly guessed, also becoming a lack of maintenance system.

Big Data Analytics for Predictive Maintenance Strategies by Lee, Cao, and Ng (2017) explained how big data application, especially to polish a corrective and preventive maintenance system condition based. Both method seems not effective and efficient, therefore to achieve a better stage of effective and efficient, within the new technology development of Big Data, predictive maintenance becoming a new breakthrough. Writer tying to explain a diagnostic and prognostic method where both can help a decision system. Diagnostic will collect every notice from the available data, while Prognostic will estimate the remaining lifetime of machine also the possibility symptoms.

An article by Dimitris Apostolou and Gregoris Mentzas "A proactive decision making framework for condition-based maintenance", discussing about a literature study development in a decision making cycle for condition based maintenance (CBM), where they proposed a framework for a proactive decision making based on condition based maintenance. The base of this research or the reference of DSS development is trying to create a prediction model based on the analysis output and data processing which coming from the real-time censors. They are trying to do a prediction based on the Remaining Use-full Life (RUL) that in the end, it will produce a DSS framework for CBM.

Maintenance strategy selection and its impact in maintenance function: A conceptual framework by Velmurugan dan Dhingra (2015) is trying to design a conceptual view from every strategic

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

maintenance which covering a corrective, preventive and predictive. From those three common strategy, writer trying to dig in more about the plus and minus and also trying to give an advice related with the best approach to implement every strategy.

Remaining Useful Life Prediction for Lithium-Ion Batteries Based on Gaussian Processes Mixture by Li, Wang, Chao, Zhou, and Xie (2016) is discussing about RUL prediction from a common lithium-ion battery used by many industry where the battery failure could create high and unexpected risk. Writer trying to explain the research stages form the design, algorithm selection and also features to be used for designing a good and accurate RUL.

According to the previous research, this research is going to discuss a downtime prediction from heavy equipment usage which is excavator. There are two data source, IoT devices and ERP system, those different data set will be connected to create a feature and target variable which can be feed-in into a machine learning system where its going to be a framework based on the predictive maintenance to produce condition based maintenance system. By using a regression method and linier regression algorithm, Support Vector Regression, Artificial Neural Network to get the best model. Along with the generated models, the accuracy evaluation of the prediction outcome based from Mean Absolute Deviation and average of Mean Square Error. After the evaluation is complete, then choose the highest accuracy score so that could become a final decision model.

3. RESEARCH MODEL

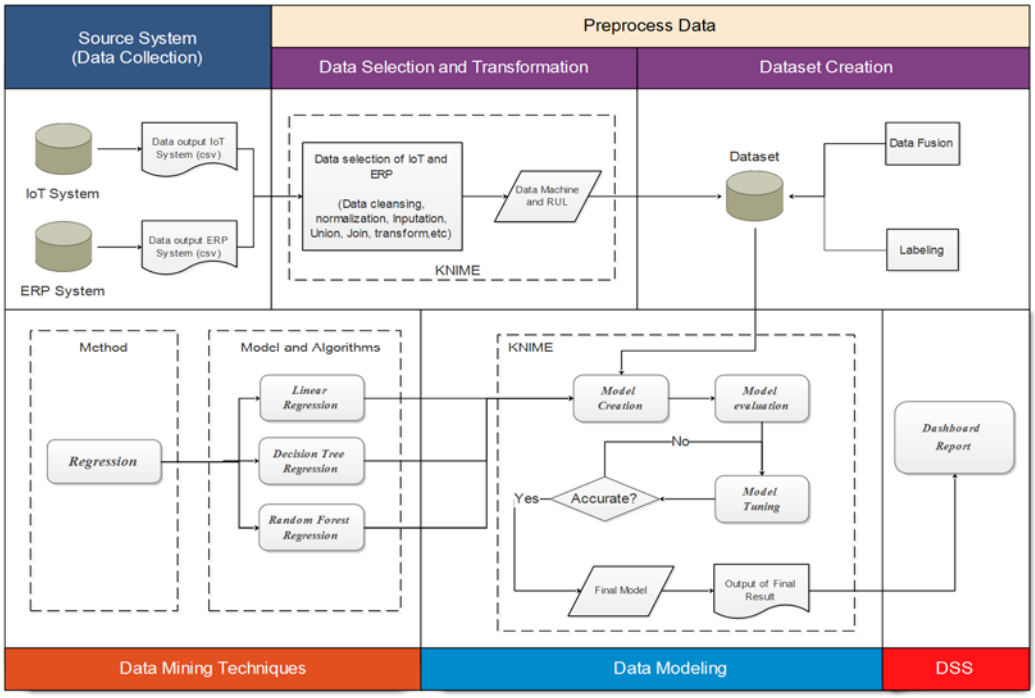

Figure 1: Research Model

3.1.1 Research Methodology

A. Data Collection

During the data collection process, to be understand that data collected from 2 different sources, data collected from ERP system DBR SAP which contain historical data of downtime, and from IoT system which contain data log from the existing IoT devices.

From those two different data set will be combine to create a feature and variable target which can be feed-in into Machine Learning system.

B. Preprocess Data

In fact, the data coming from different sources which are not directly correlated. Need to add the steps to connecting these data becoming one integrated table tha can be ready to be consumed by

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the machine learning algorithm. On this stage, correlation between these data is build based on the data schema.

Data normalization based on the tables which going to be use by dividing the data into each column with a maximum score. Normalization needs to performance in order to prevent a domination between columns.

Imputation, concerning the amount of the invalid or empty data, but there are some limit of invalid number of data with 20&% at max. Therefore if the invalid data more than 20% then we will perform data cleansing, and those column or feature will be eliminated. The use of similar imputation will be implement to the train and testing dataset. Besides from imputation, every data classification also need to binary function because of machine learning model can only accept a numeric data.

Two important things to connecting the data are the "SMR" and timestamp breakdown.

SMR is equivalent with operational machine hour which its data collected every 20 SMR, so data will be pulled off into the system every 20 hour. Where the other data contain the time details of machine failure or stopped because of damage.

C. Data Splitting

The most important part of machine learning is to validate the performance which have been built. During the validation, we do need the real data after validation as a comparison with the predicted outcome from the machine learning.

Therefore , we are using the last data breakdown to do validation, so we can see the graphic below:

Figure 2: Data Spliting

By dividing the data above, the regression function can be compare on each point from the degression of RUL, so the error rate between prediction and actual can be clearly seen.

Feature selection is a process among data observation, to find variables for the attributes and deciding which are the usable data, by separating them with the non usable data for the research such as noise or irrelevant, (Berral et al.,2012)

D. Data Training and Data Test

After the attribute and label is chosen, next step is to do train on the model. To do that, sample data have to be good, which mean allowed to have a minimum noise and double score.

According to *Blum AL (Blum & Langley, 1997),* who gave an example on how the *n feature* or *n attribute* used to indicating sample and each of *i* feature inside domain *Fi.* Explained that a feature can be a *Boolean (is_red?)*, or discrete with a high score (*discrete with multiple value (what_color?)*, or continuous *(what_wavelength?)*. While sample is a point of *instance space F1 × F2 × … × Fn. algorithm learning* given as a series of S training data, where every data point is a sample paired with *match label or classification (associated label or classification),* can be a *Boolean, multiple value or continuous.*

After the data is ready, and data model is trained, the next step is to do score prediction from the testing-dataset. This can include some of iteration adjusting with the *hyper-parameter* or *feature selection.*

E. Cross Validation

To predict the variables of the linier regression model, can be done by taking some various sample from training data repeatedly, then do the fitting against the regression linear model for every new sample and do verification on which variables could produce a better model. This approach could give us some information which are not available if the model using the actual training data and do the sampling at one time. (James, Witten, Hastie, & Tibshirani, 2013).

Cross validation method is suitable to predict an error score related with the statistic methodology to do performance evaluation, or to sort out the appropriate level of flexibility. The performance evaluation model known as the model assessment, while the process to select the appropriate level of flexibility for a model known as model selection.

When we estimating the error test against the fitting from a model which using statistic methodology, this validation is randomly dividing the observation collection into two part, the first one will be used as training set and the other one as validation set or hold-out set. This model do the fitting on training data, after that the model will be used to predict responses using data validation. Usually the outcome of the validation using MSE (Mean Square Error) for quantitative subject - produce the level or error testing. And this research is using k-fold cross validation methodology..

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

F. Data Processing

On the data processing step, also cover the design of some machine learning algorithm that will be used to do data prediction. And this stage, the tools that will be used to do data processing is KNIME. This tools will do the data analysis, modeling and data visualization.

G. Data Visualization

The purpose of data visualization in this research is to provide an accurate and effective data through graphic presentation. By having data visualization this will simplify the researcher to understand the data compare with understand a raw data. Some data visualization approach can be used, such as Box plot, Scatter Plot, Line plot , Principal Component Analysis. This research is using Line Plot visualization.

Line plot is a diagram that presentating data with lines. Whether it's a straight line, curve or dash line. This diagram can be used to present a statistic data which obtained through an observation from time to time sequentially. The common implementation is using Y and X axis. the X axis used to point the observation periode, while Y axis issued to point the outcome of the observation core during periode of time. Time collection and the observation outcome creating dots in between X and Y axis, then connect every column from contiguous dots using straight line so it will create a graphic diagram so called line graphic.

H. Evaluation

Evaluation stage is a stage where we do testing against accuration outcome from the forecast on each model. According to Heizer and Render (2009:145), there are some common calculation to count the total of forecast error. This calculation can be used to compare some different forecast model, also to oversee and ensure the prediction going well. Some indicators will be used to validating the prediction model. The common use indicators is Mean Absolute Deviation, and Mean Square Error. On the evaluation stage of this research, it will use the Mean Absolute Deviation parameter and Mean Square Error with expectation is produce the best model.

4. RESULT AND DISCUSSION

4.1 Pre-process Data

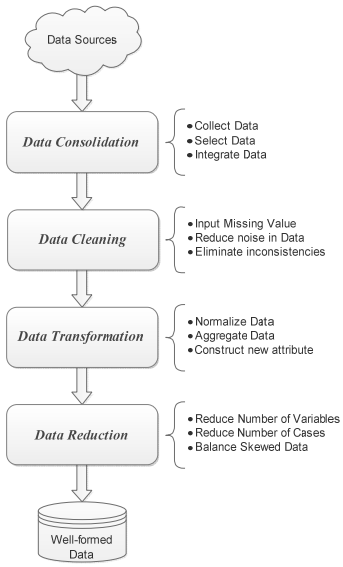

Figure 3 Preprocess Steps

Based on the figure number 3, there are 4 steps of pre-process data as mentioned below :

- First step is Data Consolidation, covered the data collection, data selection and data integration.
- Second step is Data Cleansing, covered input missing value, reduce noise in data, and eliminate inconsistencies data.
- Third step is Data Transformation, covered the data normalization, data aggregration and construction of new attribute.
- Fourth step is Data Reduction, covered reduce number of variables, cases and balancing a skewed data.

Needs to be known is that in this reserach, it only reached the third stage, without any data reduction process.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.1.2 Data Set

Table 1: Data Set

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	...	RUL
20354	20.5	1	1976	965	1.51	603	600	573	591	101	...	488.5
20354	40.8	2	1991	1154	1.87	605	618	610	621	114	...	468.2
20354	60.8	3	1992	1761	1.71	615	631	626	630	117	...	448.2
20354	80.9	4	2001	1699	1.59	611	631	620	629	123	...	428.1
20354	241.3	5	1996	1714	1.77	622	648	646	634	122	...	267.7
20354	401.5	6	2001	1722	1.79	624	630	620	630	126	...	107.5
20354	481.6	7	1999	1750	1.81	639	632	620	640	126	...	27.3
20354	501.7	8	1997	1682	1.81	636	636	621	640	127	...	7.3
20354	581.8	9	2000	1725	1.9	627	635	617	633	129	...	5561.2
20354	601.9	10	1999	1724	2.2	632	661	617	641	130	...	5541.1
20354	662.0	11	2004	1729	2.06	630	778	617	638	132	...	5481.0
20354	722.1	12	2002	1732	1.99	631	793	620	636	130	...	5420.9
20354	762.2	13	1997	1735	2.0	640	756	629	644	129	...	5380.8
20354	782.2	14	2001	1705	2.02	638	814	624	642	130	...	5360.8
20354	802.3	15	2000	1670	2.04	627	772	614	625	130	...	5340.7
20354	822.3	16	2002	1744	2.12	645	797	628	646	129	...	5320.7
20354	842.3	17	2003	1776	2.04	638	896	625	643	129	...	5300.7
20354	862.4	18	2002	1748	2.16	633	794	617	637	129	...	5280.6
20354	882.4	19	2003	1769	1.92	639	785	624	634	130	...	5260.6
...	...	...	...	...	...	...	...	...	...	...	...	...
20470	18752.8	400	1992	1551	2.67	910	902	969	835	114	...	22.5

4.1.3 Data Cleaning

By using missing value tools on KNIME, the cleaning process is covering elimination of duplication data, verification of inconsistencies data and data correction. Imputation, knowing there are so many empty or invalid data, but there is a limit for the invalid data which is 20% max. Therefore the data cleansing will only performed if the percentage of invalid data is more than 20%, then its column will be eliminated. The same imputation function will be implement for the train and test dataset. Besides the imputation, every data category has to be a binary function because the machine learning models can only accept a numeric data

4.1.4 Data Transformation

Data normalization based on the table that will be used (Table 1) by dividing each column data with the maximum score on each column. Normalization is required to prevent domination column, because the numeric number is higher than the other column. Z-score normalization is used without knowing the maximum and minimum score of the normalization, and RUL is the target.

4.2 Model Creation

4.2.1 Splitting Data Training and Data Test

Making a prediction model, analysis of data processing begin with creating data splitting which dividing the original data into two part, training and testing data. Data splitting is divide or split data collection into training and testing data. The ratio of the training against testing data is 80:20.

4.2.2 Cross Validation

To cope with overfitting or under-fitting on this research, we will randomly divide the data using cross validation method by examining the model into kfold which is cv=10.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.2.3 Dimensional Reduction

The dimensional reduction performed with criterias from various data which explained by some of the first major components.

Here is the result from Dimensional Reduction with PCA method on each algorithm.

Table 2: Best Dimension After Reduce

Algorithm	Best Dimension	Training Accuracy		Test Accuracy	
		MAE	RMSE	MAE	RMSE
Linear Regression	61	859,35	1.238,72	772,67	1.152,64
Random Forrest Regression	15	987,98	1.391,97	814,84	1.208,05
Decision Tree Regression	61	1.089,10	1.855,16	982,51	1.743,56

Based on the table 2, described that, before the dimensional reduction perform there are 76 variables and after the dimensional performed using PCA for each of the algorithm, obtain the outcome where each of algorithm is having its own best dimension which Linier Regression is having 61 dimension, Random Forrest with 15 dimension and Decision Tree Regression with 61 dimension..

4.2.4 Cross Validation Result

Here is the comparison table for three different variation score error rate from cross validation value of the training data against Linear Regression Model.

Table 3 Score of Cross Validation Linear Regression Model Data Training

Error Type	cv=10
Score Error (RMSE)	1.273,9
Score Error (MAE)	872,4

Below is the comparison table for three different variation score error rate from cross validation value of the training data against Random Forrest Regression Model.

Table 4 Score Cross Validation Random Forrest Model Data Training

Error Type	cv=10
Score Error (RMSE)	1.523,3
Score Error (MAE)	1.123,7

Below is the comparison table for three different variation score error rate from cross validation value of the training data against Tree Regression Model.

Table 5: Score of Cross Validation Tree Regression Model Data Training

Error Type	cv=10
Score Error (RMSE)	1.895,4
Score Error (MAE)	1.198,5

4.2.5 Hyper-parameter Tuning Result

A. Linear Regression

Table 6: Result of Hyper-parameter Tuning Linear Regression Model

Hyper-parameter Unavailable	Desc	Training Accuracy	Test Accuracy
-	MAE	859,35	772,67
	RMSE	1.238,72	1.152,64

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

B. Decision Tree

Table 7: Result of Hyper-parameter Tuning Decision Tree Model

Hyper-parameter			Desc	Training Accuracy	Test Accuracy
Depth	Node	Child			
11	11	4	MAE	1.156,95	1.001,37
			RMSE	1.875,70	1.582,22
13	10	2	MAE	1.145,05	1.001,88
			RMSE	1.935,88	1.684,92
13	13	4	MAE	1.207,35	1.000,57
			RMSE	1.941,71	1.641,46
14	13	4	MAE	1.157,63	1.000,20
			RMSE	1.915.55	1.640,25
15	5	2	MAE	1.159.44	1.001,41
			RMSE	1.954.11	1.696,08

C. Random Forest

Table 8: Result of Hyper-parameter Tuning Random Forest Model

Hyper-parameter			Desc	Training Accuracy	Test Accuracy
Models	Node	Child			
110	4	2	MAE	1.068,56	892,09
			RMSE	1.468,99	1.263,32
110	3	1	MAE	1.054,62	892,62
			RMSE	1.439,98	1.266,65
110	2	1	MAE	1.064,29	892,93
			RMSE	1.448,38	1.267,87
105	3	1	MAE	1.045,51	893,25
			RMSE	1.431,44	1.269,20
110	4	1	MAE	1.064,89	893,52
			RMSE	1.470,56	1.268,54

Based on table 6, 7, and 8, can be known before and after outcome of the hyper-parameter tuning for every algorithm model, there is an interested trend, that every dimension have a better testing accuracy compare with testing accuracy, which is tend to become an anomaly because mostly the training accuracy is better than testing accuracy.

There are not many literature pointing the reason for this matter, but some of them showing that the lack of less random train-set split to the data could be the reason why this happen.

4.2.6 Prediction Result

Here is the comparison between prediction performance for each algorithm;

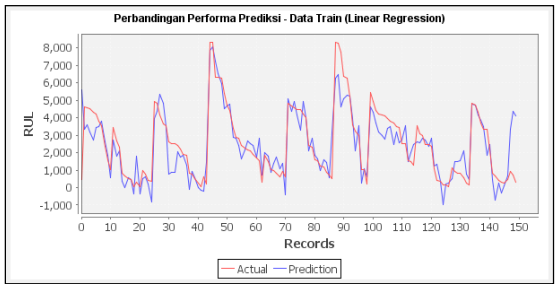

Figure 4 Comparison of Performance Prediction– Data Train (Linear Regression)

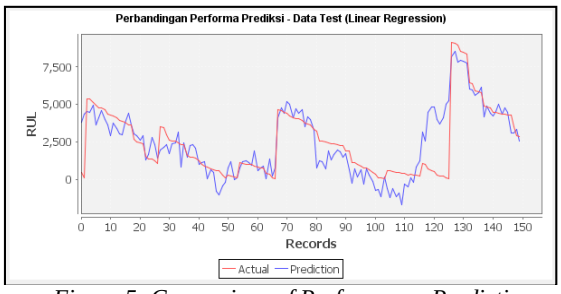

Figure 5: Comparison of Performance Prediction – Data Test (Linear Regression)

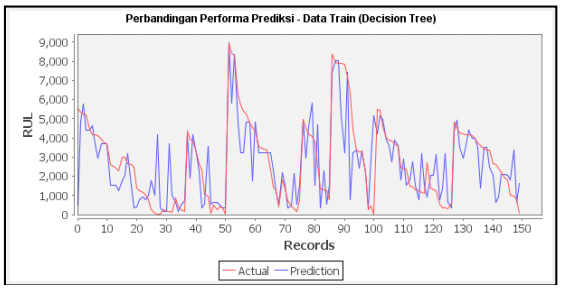

Figure 6: Comparison of Performance Prediction – Data Train (Decision Tree)

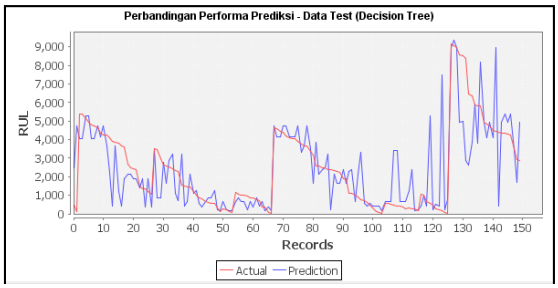

Figure 7: Comparison of Performance Prediction –Data Test (Decision Tree)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

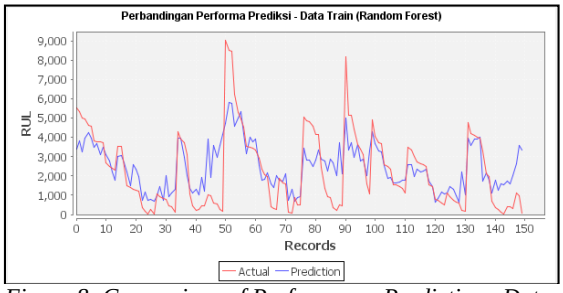

Figure 8: Comparison of Performance Prediction –Data Train (Random Forest)

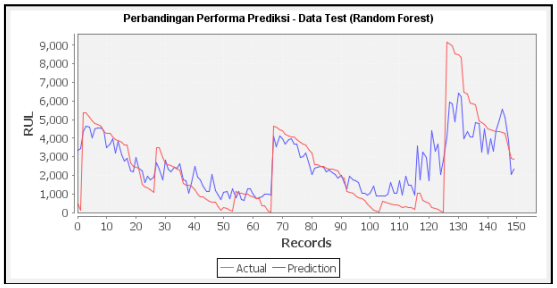

Figure 9: Comparison of Performance Prediction –Data Test (Random Forest)

Based on figure 4-9, it showing that every model has its own performance to do a prediction, and also showing the comparison of trend movement is not aligned (between actual and prediction), but this doesn't mean the model can't work, the model can still working and get evaluated in order to increase a better accuracy.

4.2.7 Final Score of Models

After knowing the comparison result from each model performance, below table will show the whole comparison of machine learning method that will have a trial

Table 9: Comparison of Error Score Machine Learning - Data Training

Score Type	Method		
	Linear Regression	Decision Tree	Random Forest
Score Error (RMSE)	1.238,72	1.875,70	1.468,99

Based on table 9, known that the error rate RMSE from each model based on training data, Liner Regression is 1.238,72 and Decision Tree with 1.875,70, and 1468,99 for Random Forest. As a visualization form from the tables, it will be convert into a graphic histogram as below;

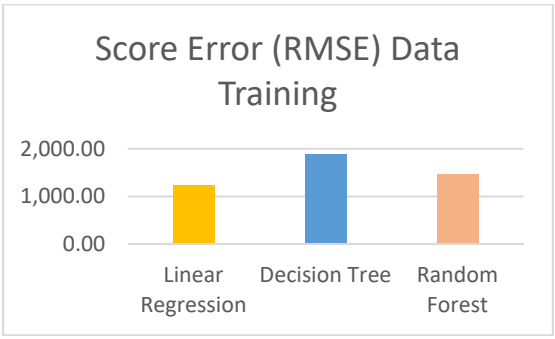

Figure 10: Comparison of Error Score Each Models– Data Training

The next step is to get the comparison of some score error (RMSE) for different machine learning model based on testing data,

Table 10: Comparison of Error Score Machine Learning - Data Test

Score Type	Method		
	Linear Regression	Decision Tree	Random Forest
Score Error (RMSE)	1.152,64	1.582,22	1.263,32

According to the table 10, known that the error rate RMSE from each of the model based on the testing data are Linear Regression is 1.152,64 and Decision Tree with 1.582,22. and 1.263,32 for Random Forest. As a visualization form from the tables, it will be convert into a graphic histogram as below :

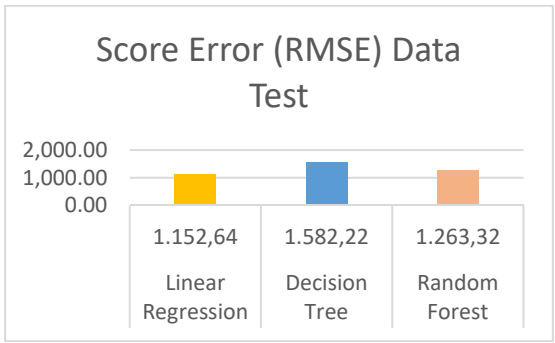

Figure 11: Comparison of Error Score Each Models – Data Test

To more clarifying the visualization for each models, it can be concluded into a single histogram graphic as below :

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

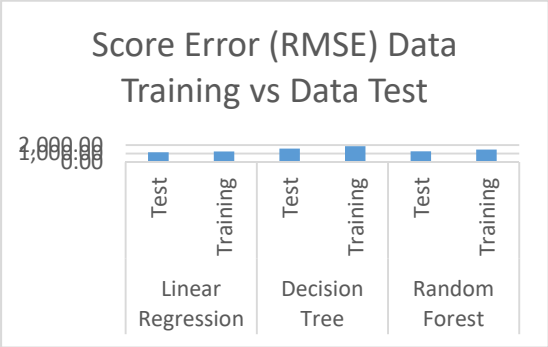

Figure 12: Comparison of Model based on RMSE –Data Training vs Data Test

After knowing the error rate comparison based on the RMSE from each models, here with the conclusion for the comparison of error rate percentage.

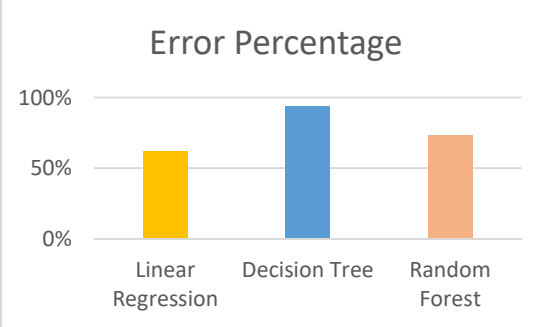

Figure 13: Comparison of Error Percentage Each Models

Based on figure 13, can be known that the error level for the Linear Regression model is 62%, Decision Tree is 94% and Random Forest is 73%. Therefor the conclusion is the lowest error level from those model is on Linier Regression with 62%.

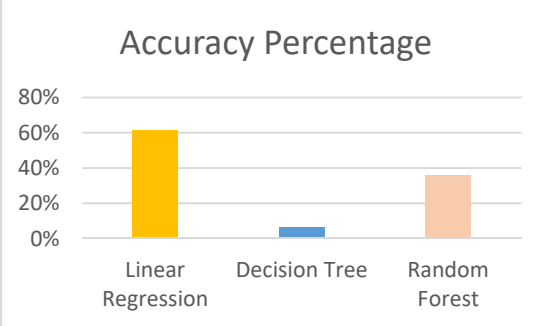

Figure 14: Comparison of Accuracy Percentage Each Models

After the error rate from the models is discovered, then we have to get the accuracy level from each models based on the error rate outcome. Based on figure 4.14, discovered that the accuracy level on Linier Regression model is 61%, Decision Tree is 7% and Random Forest with 36%. So it can be concluded if the best accuracy level from those models is Linear Regression with 61%.

4.2.8 Discussion

Referring to the explained formulation of the problem on previous chapter and based on the result of data processing above, therefore known that the available data can be used to do RUL prediction from oil grease pump spare parts by using Data Mining technique. Although the accuracy level from the data mining model is not good enough but every models have done data learning, this indicating there is still need to do evaluation against the available data source.

Working with data which have many dimension need to be very carefully and details to do its data processing. On the data processing with advance dimension, this research have done many stage, starting with data consolidation, data cleaning, data transformation, data splitting, data normalization, algorithm, cross validation, dimensional reduction until hyper-parameter tuning, those stage is a design of data processing with advance dimension in order to get the best prediction model.

According to the result of those 3 models above, Linier Regression and Random Forest model has the best performance in terms of training and testing accuracy, while decision tree is quite good but not good enough to have a good performance as random forest and linier regression, besides the Root Mean Squared Error of decision tree is bigger compare to the other two which could led to a pretty bad random impact for a prediction.

From the outcome above, the decision support system will be accommodate in order to support the business to make a decision. As the output from the KNIME is tabular version, which is quite helpful and important to the decision support system to have an interactive, friendly and effective function to be used by the users.

4.2.9 Decision Support system Result

On a stage of Decision Support System design, Passive DSS type is chosen. Passive DSS is a supporting system for decision making process, but not to give input or decision advice or any solution.

The design of Passive DSS system also referring to three basic component which is :

Journal of Theoretical and Applied Information Technology
15th July 2019. Vol.97. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Database or data source
  Output from the process of KNIME application will be the input for DSS
- Model including decision context and user criteria
  Information set as a friendly model
- User interface (UI)
  Design of the dashboard in a brief report

To develop the Passive DSS system, this research is using Microsoft Power BI. It's a Business Intelligence software, which can process a data turn to more details and present it with a good and interactive. This software can also simply visualize, control and maintaining the data.

According to the histogram on the figure 13 and 14, it showing that lowest level fault and highest accuracy level is Linear Regression model. Therefore the DSS visualization will refer to this Linier Regression model.

Herewith the output from DSS design based on Linier Regression model using Power BI tools,

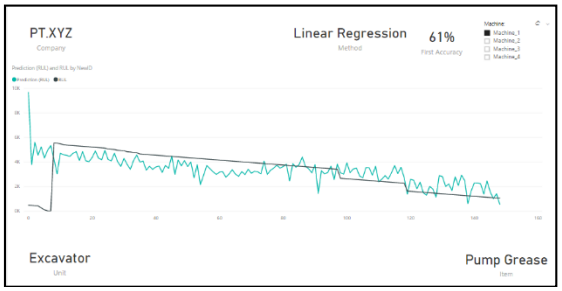

Figure 15: DSS Dashboard-1

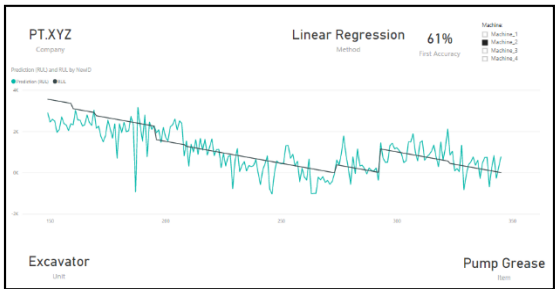

Figure 16: DSS Dashboard-2

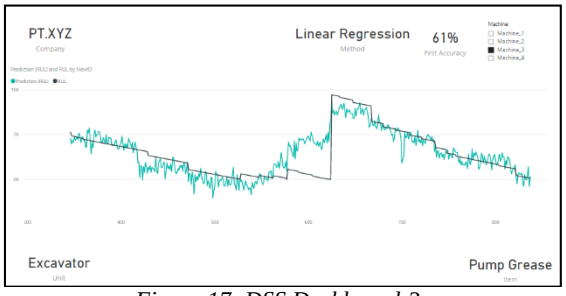

Figure 17: DSS Dashboard-3

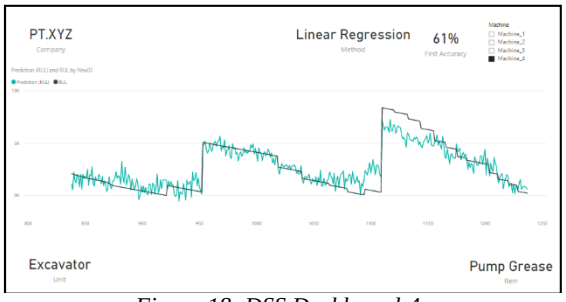

Figure 18: DSS Dashboard-4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. CONCLUSION

This research has answered its purpose based on the problem formulation explained from the beginning chapter, which is a design of Decision Support System for the decision maker. Decision Support System is designed based on the data processing with Data Mining technique, where it got some algorithm which is used to produce a comparison between the actual prediction with actual condition.

In the use of the Data Mining, this research was examining the model with cross validation steps to estimate the error score (RMSE) which related with the model that going to be implemented. But this research is doing a dimensional reduction steps before move into cross validation stage for each model, this steps can be done through a criteria of various data percentage that explained by the first main component. Dimensional Reduction also have a purpose, which is tog get the best dimension or a dimension who has a significant impact from the model.

After the best dimension from each model is obtained, the next steps is to do cross validation where each model have been through a result correction from cross validation. The testing of cross validation performed for every models which are Linier Regression, Decision Tree Regression and Random Forest Regression.

The result shows :

- The available data is not good enough to do a prediction, this can be seen from the prediction outcome which is very low for every algorithm, and there are some hypothesis for this like lack of incomplete data, lack of data distribution and the amount of the data which is very less to get a prediction schema from the Machine Learning.
- Although the available data is not good enough to do prediction, but the data still can be used to do a prediction with a matched output as the actual data condition.
- The result of comparing machine learning or algorithm method, show that the best score produced by Linear Regression model with 61% percentage for the prediction accuracy level. Which mean that the Linier Regression model can be used as prediction model for condition based maintenance decision..
- By knowing the outcome from the prediction performance, therefore decide the breakdown time from the used spare parts.
- Decision Support System as a visualization of an interactive and user friendly dashboard, so that the decision makers can quickly understand the business needs..

REFERENCES

[1] AL-Hassan, K., Chan, J. F.-L., & Metcalfe, A. V. The role of total productive maintenance in business excellence. *Total Quality Management, 2000. 11*(4-6), 596-601.

[2] Bernard, W. T. *Introduction to Management Science: Sains Manajemen* (8 ed.), 2005. (V. Silvira, Trans.) Jakarta: Salemba Empat.

[3] Campo, K., Gijsbrechts, E., & Nisol, P. The impact of stock-outs on whether, how much and what to buy,2000. *RUG*.

[4] Chandra, A. Kontribusi Industri Manufaktur Terhadap PDB. Yogyakarta, Indonesia: Medium . January 17, 2018.

[5] Chopra, S., & Meindl, P. *Supply Chain Management: strategy, planning, and operation* (3rd ed.), 2007. New Jersey: Pearson.

[6] David, F. R. *Strategic Management: Concept and Cases* (13 ed.), 2011. New Jersey: Pearson Prentice Hall.

[7] Fuller, D., Buote, R., & Stanley, K. A glossary for big data in population and public health: discussion and commentary on terminology and research methods. *Journal of Epidemiology and Community Health, 71*(11), 2017, 1113-1117. doi:10.1136/jech-2017-209608

[8] Georgakopoulos, D., & Jayaraman, P. P. Internet of things: from internet scale sensing to smart services. *Computing. Archives for Informatics and Numerical Computation, 98*(10), 2016, 1041-1058. doi:10.1007/s00607-016-0510-0

[9] Gross, A. C., & Weiss, D. D. Industry corner: The global demand for heavy construction equipment. *Business Economics,1996. 31*(3), 54.

[10] Kotler, P., & Armstrong, G. *Principles of Marketing* (14E ed.), 2012. New Jersey: Pearson Education.

[11] Kowalkowski, C., Kindström, D., & Brehmer, P.-O. Managing industrial service offerings in global business markets. *The Journal of Business & Industrial Marketing, 26*(3),2011. 181-192.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[11] Laudon, K. C., & Laudon, J. P. *Management Information Systems: Managing The Digital Firm* (12 ed.),2012. Pearson Prentice Hall.

[12] Li, L., Wang, P., Chao, K.-H., Zhou, Y., & Xie, Y. Remaining Useful Life Prediction for Lithium-Ion Batteries Based on Gaussian Processes Mixture. *PLoS One, 11*(9), 2016, 1-13. doi:10.1371/journal.pone.0163004

[13] Liker, J. K. *The Toyota Way: 14 Management Principles from the World's Greatest Manufacturer, 2004.* New York: McGraw-Hill.

[14] Pheng, L. S., Shang, G., & Peter, L. K.. Using Lean Principles to Reduce Wastes in the Concerting Supply Chain. *International Journal of Construction Project Management, 8*(1), 2016. 3-23.

[15] Rayport, J. F., & Jaworski, B. J. *Introduction to e-Commerce* (International Edition 2nd ed.),2003. McGraw-Hill/Irwin.

[16] Robbins, S. R., & Coulter, M. *Management* (11 ed.), 2012. New Jersey: Pearson Prentice Hall.

[17] Sarwono, J. *Metode Penelitian Kuantitatif dan Kualitatif* (1st ed.), 2006. Yogyakarta: Graha Ilmu.

[18] Strauss, J., & Frost, R. *E-Marketing* (6 ed.), 2012. United States: Pearson Education.

[19] Velmurugan, R. S., & Dhingra, T. Maintenance strategy selection and its impact in maintenance function: A conceptual framework. *International Journal of Operations & Production Management, 35*(12), 2015. 1622-1661.

[20] Mubarok, Muhammad Ilham Pohon Regresi dengan Pendekatan Generalized Unbiased Interaction Detection Estimation (Guide) untuk Data Multirespon, 2018. Yogyakarta, Indonesia