# Integrationof Data Mining and Design of Experiment to Diagnosis Machine Efficiency

***Hamza Saad****
*Ph.D Candidate of Engineering*
*Department of Industrial and System Engineering,Binghamton University, Binghamton, New York*
****Email:*** *hsaad1@binghamton.edu*

***DOI:*** *http://doi.org/10.5281/zenodo.2535973*

## *Abstract*

*The study focused on three phase induction motor because it is playing a vital role in production development. The motor may get the problem in some parts may effect in the wholeefficiency. Biannually maintenance increases the life of the machine, but many companies, especially in the third nation, do the maintenance as soon as the machine gets fail. The studyapplied two strategies to diagnosis and improve the efficiency of the machine. Design of experiment conducted to extract the main effect and interactions between variables. Three variables current, voltage, and power factor applied to understand the main effect in the exploited power (P). Power factor recorded the most important factor impacts the exploited power. Then, data mining utilized with two algorithms; random forests and linear regression. These algorithms used to predict the power factor — six variables collected for data mining, current, power losses, voltage, apparent power, resistance, and exploited power. These variables used to predict the power factor. From both strategies, we found that there is a strong relationship between exploited power and power factor, and those variables have a positive impact on the efficiency of the machine. Also, power losses and current have a negative impact on the power factor. Voltage did not give significant important whether in the design of experiment or data mining. Therefore, power losses and current must be controlled to keep the efficiency, and this can be done by regular maintenance by professional workers.*

***Keywords****: Design of experiments, Data mining, Feature selection, Efficiency, Three-Phase Induction motor*

## INTRODUCTION

Induction motors are used for converting electric energy into mechanical energy. They utilized for daily operations — the loads on the machine whether low, medium or high are conducted by single and three phase induction motors. The induction motor iscomprehensively applied in washing machines, mixers, dryers, air conditioners, refrigerators, and fans [2]; in addition to textile machines [1]. Machines now have a significant improvement in construction and performance. Using excellent conductivity materials like copper as an alternative to the standard steels in the squirrel cage rotor decreases power losses in the motor. Also, the silicon steel sheets use to shape the stator core decreases the hysteresis losses and eddy currents [4].

The speed of three-phaseinduction motors can be controlled by changing the used voltage to the stator, some stator poles, stator frequency or by connecting an external resistor to the rotor circuit (and this controlling in wound rotor type) [5]. Variable frequency drives are commonly utilized to control the speed of induction motorstosave energy in the motor. For example, throttles devices, valves or dampers can be applied to adjust the output of water pumps to decrease the flow rate, but this increases the current drawn by the pump. Thus, variable frequency devices are utilized to control flow by varying motor speed. Changing the frequency and voltage are applied to the pump instead of adjusting flow through throttling devices. Therefore, the energy saving is accomplished. Different techniques such as step down auto-transformer, star/delta connection and adding resistors in series with the rotor (slip ring rotor) or stator windings areutilized to reduce starting current of induction motors, which is 5 to 7 times its rated current [6].

Three faze induction motorist utilized in most industrial fields due to many advantages, like strong structure, high efficiency at rated speed, low weight, easy maintenance and affordable cost [7]. However, the disadvantages of the three-phase induction motor aredifficult to control in the rotor speed, at light load the efficiency and power factor are low, and high starting current [7].

Improving the performance and efficiency of three-phaseinduction motors, which is the aim of this study; plays a significant role in decreasing the energy wastes. Different strategies have been applied for improving and diagnosing the performance of induction machines.

## METHODOLOGY OF STUDY

The study is for improving the performance of the machine by collecting most of the variables which impact the machine performance. The machine has two problems might occur whether from inside or outside. For outside problems, the problem may come from bad maintenance, or bad operating.

Moreover, the inside problem may come because internal fails and this what the study will focus on — two strategies applied to get a full understanding of the three-phase induction motor; data mining and design of experiments. In data mining, two algorithms applied to predict the power factor; these algorithms are random forests and linear regression. Both algorithms give well prediction and understanding of the data. In the design of the experiment, the central composite (Response Surface) will apply to extract the main effect and interactions on the exploited power. Output in both strategies has switched, in DOE power factor has used as an input variable and exploited power (P) has used as response and in data mining power factor has chosen as output and exploited power (P) has chosen as an input variable. Both variables power factor and exploited power (P) are much related together; this means the improvement in exploited power (P) it leads to improvements in power factor and whole machine performance. In data mining, only power factor gave a very small error.

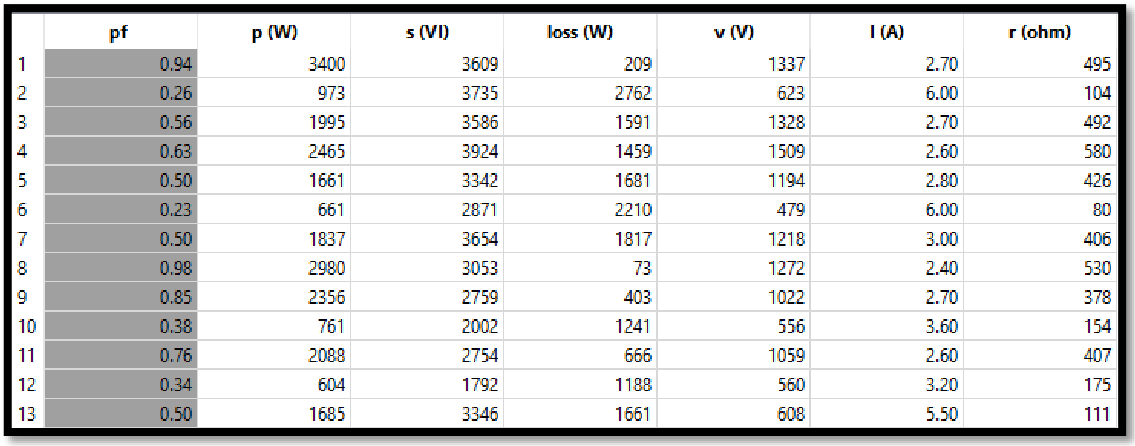
The output of DOE and data mining are not the same for different reasons;

1. Power factor and exploited power (P) have a positive impact on the performance of the machine.
2. Exploited power (P) works well in DOE, but it gives high errors in data mining, thus bad prediction obtained if it applied as output in data mining.
3. Data in data mining has increased to cover all variables that may affect the power factor.
4. Variablewhen works as an input and output givemore details, as instance power factor works in DOE as input and data mining works as a dependent.

## DATA COLLECTION

Data is collected based on two strategies. First, data for the design of experiments, few variables selected to understand the exploited power (P) that applied as a response in DOE. Three variables are power factor, current, and voltage modified to predict and understand the impact on exploited power (P). Second, data for data mining, in this case, we increased the data to apply data mining to extract the important variables which have a high impact on the power factor. Six variables applied in data mining to predict power factor these variables are; exploited power, apparent power, power losses, voltage, current, and resistance. Data of design of the experimentis presented in Table 2, and data of data mining is presented in Table 1, where power factor is applied as dependent variable and rest of the variables as predictors to the dependent variable. Data in Table 1 has collected by testing real machine, but other variables like Losses and apparent power collected using the electrical equation of power system in [8].

***Table 1:*** *Part of Dataset.*



| | pf | p (W) | s (VI) | loss (W) | v (V) | I (A) | r (ohm) |
|---|---|---|---|---|---|---|---|
| 1 | 0.94 | 3400 | 3609 | 209 | 1337 | 2.70 | 495 |
| 2 | 0.26 | 973 | 3735 | 2762 | 623 | 6.00 | 104 |
| 3 | 0.56 | 1995 | 3586 | 1591 | 1328 | 2.70 | 492 |
| 4 | 0.63 | 2465 | 3924 | 1459 | 1509 | 2.60 | 580 |
| 5 | 0.50 | 1661 | 3342 | 1681 | 1194 | 2.80 | 426 |
| 6 | 0.23 | 661 | 2871 | 2210 | 479 | 6.00 | 80 |
| 7 | 0.50 | 1837 | 3654 | 1817 | 1218 | 3.00 | 406 |
| 8 | 0.98 | 2980 | 3053 | 73 | 1272 | 2.40 | 530 |
| 9 | 0.85 | 2356 | 2759 | 403 | 1022 | 2.70 | 378 |
| 10 | 0.38 | 761 | 2002 | 1241 | 556 | 3.60 | 154 |
| 11 | 0.76 | 2088 | 2754 | 666 | 1059 | 2.60 | 407 |
| 12 | 0.34 | 604 | 1792 | 1188 | 560 | 3.20 | 175 |
| 13 | 0.50 | 1685 | 3346 | 1661 | 608 | 5.50 | 111 |

## DESIGN OF EXPERIMENTS

Central composite design is useful in the methodology of the response surface to build a second order (quadratic) model for the response without needing to apply a complete experiment of three-level factorial. As soon as the designed experimentis conducted, linear regression is applied, sometimes iteratively to get results. Coded variables are usually used when establishing this type of design.

In design, three factors have selected to understand the main effect in the exploited power (P), Power factor value is 0.3 as low level, and 0.95 as high level, the current value (I) is 2.4A as low level and 6A as high level, and voltage (V) is 657V as low level and 675V as high level.

***Table 2:*** *Design of Experiments. Central composite (Response Surface)*

| 3 Input factors with 2 levels | | | Response |
|---|---|---|---|
| **Power Factor (PF)** | **Current (I)** | **Voltage (V)** | **Exploited power (P)** |
| 0.3 | 2.4 | 657 | 473.04 |
| 0.3 | 6 | 657 | 1182.6 |
| 0.3 | 6 | 675 | 1215 |
| 0.95 | 2.4 | 657 | 1497.96 |
| 0.95 | 6 | 657 | 3744.9 |
| 0.95 | 6 | 675 | 3847.5 |
| 0.95 | 2.4 | 675 | 1539 |
| 0.3 | 2.4 | 675 | 486 |
| 0.3 | 6 | 675 | 1215 |
| 0.95 | 2.4 | 657 | 1497.96 |
| 0.95 | 6 | 657 | 3744.9 |
| 0.95 | 6 | 675 | 3847.5 |
| 0.95 | 2.4 | 675 | 1539 |

Preparing and executing the design of the experiment is not pretty easy. We measured and tested the output of the machine and confirmed all data using the electrical equation in [8].

Design of experiments built for three factors and each factor includes two levels.

***Table 3:*** *Main Effects Results.*

| Factor | Effect | St. Err | t(6) | P |
|---|---|---|---|---|
| Mean/interact | 1747.771 | 2.140527 | 816.5146 | 0.000000 |
| (1) PF(L) | 1819.137 | 4.281054 | 424.9275 | 0.000000 |
| (2) I (L) | 1497.543 | 4.281054 | 349.8070 | 0.000000 |
| (3) V(L) | 9046.293 | 4.281054 | 10.8134 | 0.000037 |
| 1L by 2L | 780.177 | 4.281054 | 182.2395 | 0.000000 |
| 1L by 3L | 25.527 | 4.281054 | 5.9628 | 0.000996 |
| 2L by 3L | 23.122 | 4.135891 | 5.5905 | 0.001393 |

To remove additional results in research, we only showed the result of the main effect to show how factors impact on the exploited power. Factors have significantimportance,and interactions between factors have less important than main factors. Power factor is the most important variable, then current and interaction between the power factor and current respectively.

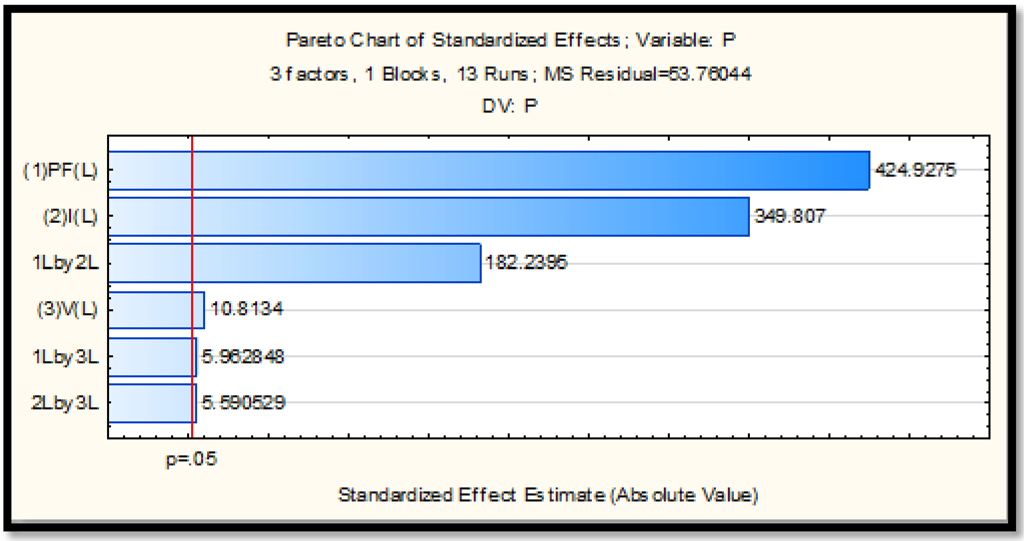


***Figure 1:*** *Pareto Chart for the Main Effects.*

In Figure 1, it appears that the power factor has the highest impact on the response. Voltage has a low impact on the response. Also, all interactions of the voltage are less significant. Only current and power factor have significantimportance as shown in Figure 1.

**Data Mining Application**

Two algorithms applied to extract the knowledge about the three-phase induction motor. Dataset if following linear relationship, and the algorithms applied for regression prediction because data is in numerical form. The algorithms are:

***Random Forests***

The algorithmis built by combining the predictions of many trees; each tree is trained in isolation [8]. Typically, the tree is trained independently, and the results of the trees are combined through the average if algorithm applied for regression and vote to the most popular result if algorithm applied for classification. Three choices need to be done when conducting a random forests tree. (1) Splitting the leaves, (2) type of predictor used in each leaf, and (3) the method for injecting randomness into the trees. Identifying the method for splitting the leaves required for selecting the shapes of candidate splits as well as a method to evaluate the quality of each candidate. The typical choices are to use the axis aligned splits, where the data are routed to the sub trees depend on whether or not they exceeded the threshold value in a linear split; or chosen dimension, where linear features combination are the threshold to decide ultimate decision. The threshold value in the algorithm either case can be randomly chosen or by optimizing the function of the data in the leaves. To split a leaf, a collection of candidate split is generated, and a criterion is evaluated to select between them. A simple method is to choose among the candidates randomly and uniformly, as in the models analyzed in [7]. More common approaches are to choose the candidate split that is optimizing a purity function over the leaves that would be constructed. Here, maximizes the information gain is the typical choice [11]. The most common option for predictors for each leaf is to apply the average response for all the training points which fall into that leaf. [9] explored the use regression and other tasks for different leaf predictors. However, these generalizations are beyond the scope of this paper study. The case study considered regression prediction because the problem will analyze based on regression. Injecting the randomness within the tree construction can happen in many ways. The choice of which dimension to be used as split candidates at each leaf can randomize, in addition to the choice of coefficients for random features combination. In either case, the thresholds can be selected by the optimization over all or some of the data in the leaf, or randomly. Another conventional technique to introduce randomness is by building each tree using a Boots trapped learning or sub-sampled data set. In this case study, each tree in the forest is trained using Bootstrap, which introduces differences between the trees based on sampling with replacement. For all dataset, random forests from Orange data mining in Figure 2 has analyzed data by constructing 15 trees to improve the final prediction [10]. The number of trees selected by the analyzer to evaluate the change in accuracy to pick the best accuracy. Fifteen trees recorded the high accuracy, then over 15 trees call over fitting problem.

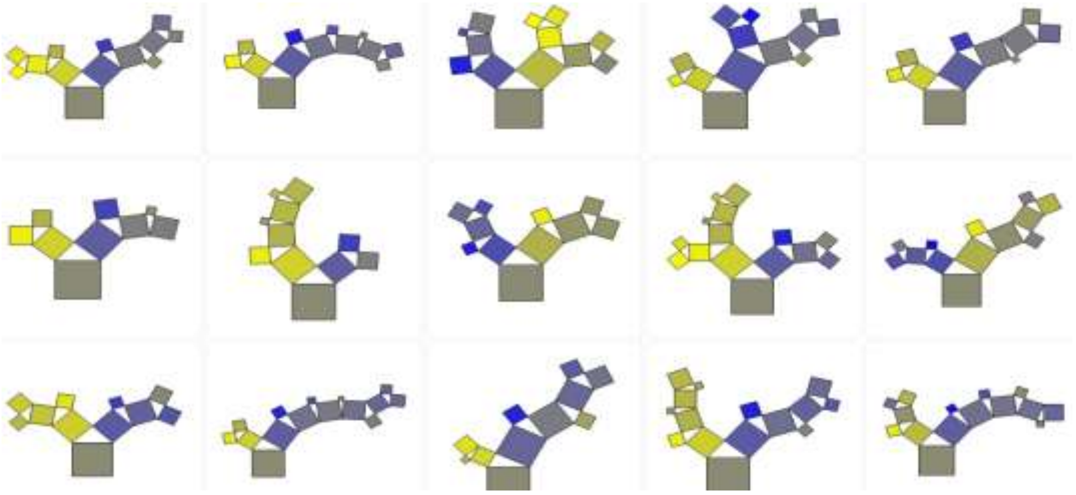


***Figure 2:*** *Random Forest Trees.*

As presented in Figure 2 for each tree, the highest value of power factor is at yellow nodes; tree construction has managed the scale of bad power factor at blue nodesand the high power factor at the yellow nodes. The value of power factor reaches over 90% at the pure yellow, and this value is very suitable to save more energy and keep the performance of motor up. However, which variables or predictors help to predict high power factor, the best attributes will present in feature selection in Table 5. For all trees, we can figure out that most of the splits for high power factor (yellow) in the left side and the split for high power factor is more comfortable than bad power factor.

***Linear Regression***

This algorithm is a common statistical analysis before applying in data mining as an independent algorithm. In statistical modeling, a regression analysis is applied to estimate the relationships between two or more factors as a dataset of this study:

The dependent variable is the primary variable used to understand and predict.

Independent variables are the variables that might impact the dependent variable.

Regression analysis helps to understand how the dependent variable can be changed when one of the independent variables changes and these changes will determine mathematically for which variable makes an impact on the dependent variable. A regression analysis model is according to the calculation of the sum of squares, which is a mathematicalapplication to find the dispersion of data points. A model aims to obtain the smallest sum of squares and draw the line that is closest to the data as possible.

In statistics, the difference between multiple linear regression and simple regression is that the relationship between a dependent variable and one independent variable in simple linear regression models is using a linear function. If two or more explanatory variables used to predict the independent variable the way it dealswith themultiple linear regression. If the dependent variables are modified as a non-linear function because the datais not followed a linear relationship, nonlinear regression is suitable for non-linear data. However, a dataset of study is followed simple linear because in table 3 the sum of square is very small and $R^2$ is very high; this means the dependent variable is much related to dependent variables.

Mathematical linear regression equation is

$y = bx + a + \varepsilon$……………….. (1)

In Figure 2 we presented how software applied to solve data in table 1. So, in table 4, random forests and linear regression gave sufficient understanding to data of three-phase induction motor Now, use these algorithms to predict power factor (PF) using predictors power losses, apparent power (S), resistance (R), current (R), voltage (V) and exploited power (P). In Figure 3, there are some icons will be explained as follows;

***Predictions*** used to calculate and measure the relationship between observations and predicted values by calculating $R^2$.

***Rank*** used to rank the mostimportant variables based on the most important variable.

***Pythagorean Forest*** is special for the random forest to show how tree constructed and nodes how to split in a simple graph.

***Pythagorean Tree*** is connected with Pythagorean Forest to give one tree by averaging 15 trees. Also, it used to show how the dependent variable split in the nodes.

***The Data Table*** isto check if there is any missed data before analyzing.

Test/Scourer to give the output of each algorithm of data mining in organized details:

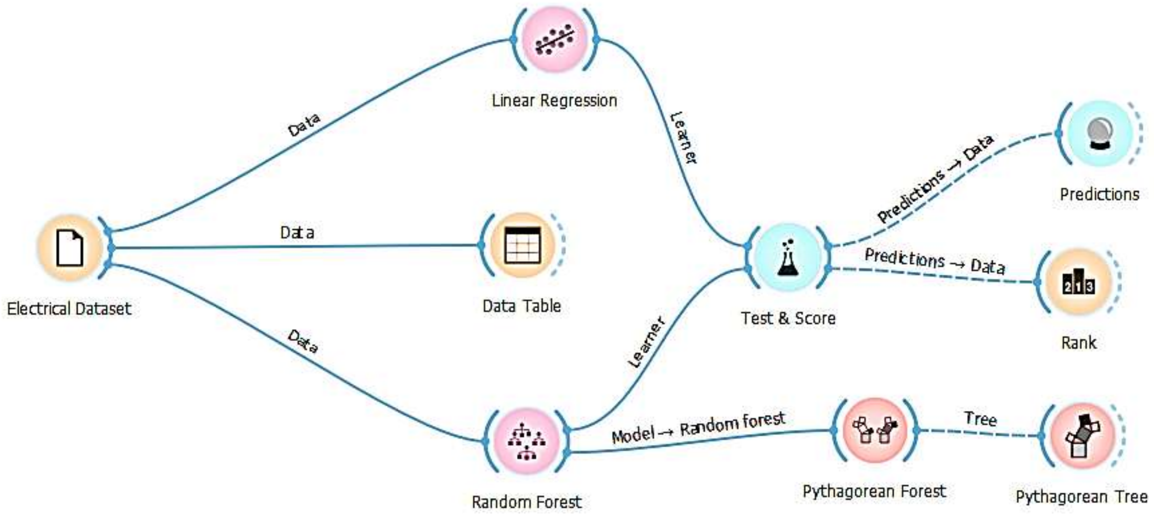


***Figure 3:*** *The Application of Data Mining to Predict Power Factor.*

Results from data mining algorithms of Figure 3 are presented in Table 4.

***Table 4:*** *The Output of Each Algorithm.*

| **Algorithm** | **MSE** | **RMSE** | **MAE** | **R²** |
|---|---|---|---|---|
| Random Forests | 0.006 | 0.078 | 0.059 | 0.873 |
| Linear Regression | 0.002 | 0.048 | 0.034 | 0.952 |

### *Ranking Variables*

All variables considered by each algorithm to predict power factor (PF), each input variable has a particular impact, and we can extract and rank each impact using feature selection. Input variables may give positive or negative effect in the output. From this rank in Table 5, power losses ranked as the first important variable impacted on the power factor, and then exploited power. However, voltage and apparent power do not give a significant impact on the power factor. Four variables may confirm to build logic decision about the performance of machine by matching results of data mining and the design of experiments together to enhance one result.

***Table 5:*** *Feature Selection*

| **Variable** | **Univariate repression** | **RReliefF** | **Rank** |
|---|---|---|---|
| Power losses | 635.059 | 0.255 | 1 |
| Exploited power (P) | 220.261 | 0.241 | 2 |
| Current(I) | 181.231 | 0.186 | 3 |
| Resistance (R) | 115.421 | 0.141 | 4 |
| Voltage (V) | 70.534 | 0.126 | 5 |
| Apparent power (S) | 0.992 | 0.195 | 6 |

## CONCLUSIONS

In the design of experiments, the power factor is much related to exploited power and voltage did not show real work to predict the exploited power. The interaction between the power factor and current is more important than voltage. All interactions of the voltage do not show significantly. Also, the current should be related to power losses especially if there is high resistance. In data mining, random forests and linear regression give high R² and very small error. Also, power losses ranked as the most important variable impacts on the power factor. We can understand why the machine has low efficiency? Because the power losses existed in the machine, and this loss supported by resistance and current. Data mining and design of experiment give the same result about power all variables such as voltage is not important, current is important, power factor and exploited power are very important in both strategies. However, power losses are the first variable effect on power factor. Also, there is a strong relationship between power losses and current, so this relationship dropped the efficiency down. Power losses and current help to predict low power factor, and in the design of experiments, the current impact from a high level on the exploited power. Maybe we did not collect all the related data, but regular maintenance should be done to remove all problems that lead to power loss or increase current.

## REFERENCES

1. H. Saad. Application of Data Mining to Improve Evaluation Performance. *Beau Bassian, Mauritius. Scholar's Press*. 2018.
2. Sharma S., Gaur B. & Punetha D. Optimization Technique to Mitigate the Losses in Single Phase Induction Motor. *International Conference on Advances in Computing, Communication, and Automation (ICACCA), Dehradun*. 2016, pp. 1-4.
3. Saad H. Use Fuzzy Rules and Experimental Design to Predict and Improve Output Performance of Three-Phase Inductive Motor. DOI: *10.15662/IJAREEIE.2017.0607003*. 2017.
4. Takahashi I., Koganezawa T., Su G. & Ohyama K. A Super High Speed PM Motor Drive System by a Quasi-Current Source Inverter. *In IEEE Transactions on Industry Applications*. 1994. 30(3), pp. 683-690.
5. Bakshi M. B. U. Transformers and Induction Machines. Technical Publications. 2009.
6. Larabee J., Pellegrino B. & Flick B. Induction motor starting methods and issues. *Record of Conference Papers Industry Applications Society 52nd Annual Petroleum and Chemical Industry Conference*. 2005. pp. 217-222.
7. Theraja B., Theraja A., Patel U., Uppal S., Panchal J., Oza B., Thakar V., Patel M. & Patel R. A Textbook of Electrical Technology: Vol II. *S.Chand Publishers*. 2005.
8. Saad H. Use Data Mining and Statistical Analysis to Improve the Efficiency of the Three-Phase Induction Motor Based on the Power Factor. DOI: *10.15662/IJAREEIE.2018.0708002*. 2018.
9. Criminisi A., Shotton J., & Konukoglu E. Decision Forests. A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. *Foundations and Trends in Computer Graphics and Vision*. 2011, 7(2-3): 81227.
10. Saad HR. Use Bagging Algorithm to Improve Prediction Accuracy for Evaluation of Worker Performances at a Production Company. *Ind Eng Manage*. 2018, 7: pp. 257. Doi: *10.4172/2169- 0316.1000257*.
11. Hastie, T., Tibshirani, R., & Friedman, J. The Elements of Statistical Learning. Springer, 10 edition. 2013.

**Cite this article as:**
Hamza Saad. (2019). Integrationof Data Mining and Design of Experiment to Diagnosis Machine Efficiency. Journals of Advancement in Mechanics, 4(1), 1–8. http://doi.org/10.5281/zenodo.2535973