Use Data Mining Techniques to Identify Parameters That Influence Generated Power in Thermal Power Plant

Waleed Hamed Ahmed Eisa [1], Naomie Bt Salim [2]

Faculty of Computer Sciences, Sudan University of Science and Technology1
Faculty of Computer Science and Information Systems, University Teknologi Malaysia[2]
mds.waleed@gmail.com

Received: 02/08/2016
Accepted: 29/10/2016

ABSTRACT- The goal of this paper is to identify the parameters that influence the amount of power generated by steam power plants. Data mining tools were used to prove that influencing parameters are differ according to the current status of power plant. Waikato environment for Knowledge analysis (Weka) was used for feature selection and building the prediction model. An initial comparison between many algorithms for each data set was reported. Then the prediction model was built using linear regression algorithm, because it shows the highest correlation coefficient between parameters, and minimum errors. The selected model predicts the generated power using all available parameters as predictors. Although this is not a practical method for power prediction, because not all predictors are controllable, but it reflects how much a parameter influence the amount of generated power. Evaluation results of these models were discussed and a detailed analysis sheet was prepared, to prove that data mining is the best way to predict the amount of generated power, and show the health status of steam power plants.

المستخلص تهدف هذه الورقة لتحديد العوامل التي تؤثر على تحديد كمية الكهرباء المنتجة من محطات التوليد الحراري. تم استخدام تقنيات التنقيب في البيانات لإثبات أن أثر هذه العوامل يختلف باختلاف حالة المحطة. بهذا البحث تم استخدام حزمة تحليل المعرفة Weka لتحديد العوامل المؤثرة و بناء نموذج التنبؤ. حيث تم أولاً إجراء مقارنة أولية بين نتائج تشغيل عدة خوارزميات على مجموعات مختلفة من البيانات, ومن ثم تم بناء نموذج التنبوء باستخدام خوارزمية الانحدار الخطي, حيث أظهرت الخوارزمية أعلى معامل ارتباط بين العوامل المختارة , و أقل نسبة خطأ. يقوم النموذج المختار بتوقع كمية الكهرباء المنتجة من المحطة باستخدام كل العوامل المتاحة بمجموعة البيانات, و مع أن هذا ليس اسلوباً عملياً , لأنه لا يمكن التحكم بكل هذه العوامل, إلا أن النموذج يعكس مدى أثر هذه العوامل على تحديد كمية الكهرباء المنتجة. أيضاً تم تقييم و مناقشة هذه النماذج, كما تم إعداد تحليل تفصيلي لأثر هذه العوامل. هذا البحث يثبت أن التنقيب في البيانات هو أفضل طريقة لتوقع كمية الكهرباء المنتجة من محطة التوليد الحراري, و دراسة حالتها الحالية.

Keywords*: Power Plant , Thermal Power Plant , Feature Selection , Prediction, Regression, Data Mining*.

INTRODUCTION

The availability of real time data in the electric power industry encourages the adoption of data mining techniques. Data mining is defined as the process of discovering patterns in data [1]. However, there are some obstacles that face researchers and engineers to benefit from data mining in this area. The first one is the interdisciplinary nature of such a research, because it requires deep knowledge in both IT and electromechanical engineering. Another obstacle is the lack of standard analysis methods and benchmarks, this leads to usage of different methods and datasets [2].

A) Data Mining

Data mining is defined as the process of discovering patterns in data [1]. To discover this pattern; first we must explore the past then we can

predict the future. Exploring the past is done by describing the data by means of statistical and visualization techniques. While future prediction is done by developing prediction models. Prediction models could be classified to the following categories [3]:

(1) Classification: if the model outcome is categorical. There are four main groups of classification algorithms:
   1. Frequency Table: which contains ZeroR, OneR, Naive Bayesian and Decision Tree algorithms.
   2. Covariance Matrix: which contains Linear Discriminant Analysis, and Logistic Regression.
   3. Similarity Functions: which contains K Nearest Neighbors
   4. Others: which contains Artificial Neural Network, and Support Vector Machine

(2) Regression: if the model outcome is numerical. There are four main groups of regression algorithms:
   1. Frequency Table: which contains Decision Tree
   2. Covariance Matrix: which contains Multiple Linear Regression
   3. Similarity Functions: which contains K Nearest Neighbors
   4. Others: which contains Artificial Neural Network, and Support Vector Machine

(3) Clustering or descriptive modeling: is the assignment of observations into clusters so that observations in the same cluster are similar. There are two main groups of clustering algorithms:
   1. Hierarchical which contains Agglomerative, and Divisive.
   2. Partitive: which contains K Means, and Self-Organizing Map.

(4) Association rules: can find interesting associations amongst observations. One of the most famous algorithms in this group is Apriori algorithm which was proposed by Agrawal and Srikant in 1994 [4]. The algorithm generates the association rules by finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum support). Then the algorithm uses the larger itemsets to generate the association rules that have confidence greater than or equal to a user specified minimum confidence [5].

B) Power System

The power system which is also known as the grid is divided into three components: the generator which produce the power, the transmission system that carries the power from the generators to the load centers, and the distribution which delivers power to the end users.

There are many types of generators (also known as power plants) normally these power plants contain one or more generators, which is a rotating machine that converts mechanical power into electrical power. Then the motion between a magnetic field and a conductor creates an electrical current. Most power plants in the world burn fossil fuels such as coal, oil, and natural gas to generate electricity. Others use nuclear power, but there is an increasing use of cleaner renewable sources such as solar, wind, wave and hydroelectric [2].

C) Rankine Cycle

This research focus on thermal power plants that uses oil as energy source, these types of power plants uses Rankine Cycle to generate power. More theoretical investigation about Rankine Cycle could be found in [6]. Rankine Cycle is a closed system consists of four main components, that are interconnected together to build one system [7]. As shown in figure 1 below, these components are:

1. **Steam Turbine which** uses the superheated steam that is coming from the boiler to rotate the turbine blades.
2. **Condenser**: uses external cooling water to condense the steam which is exhausted from turbine to liquid water.
3. **Feed water Pump**: to pump the liquid to a high pressure and bush it again to boiler.
4. **Boiler** which is externally heated to boil the water to superheated steam.

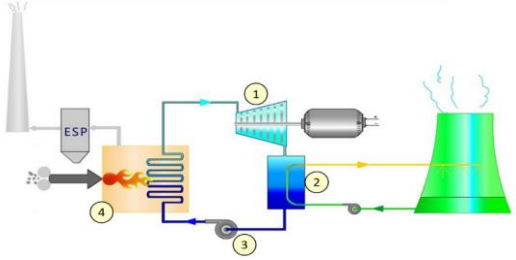

Figure 1 Thermal Power Plant using Rankine Cycle [7]

D) The use of data mining in power plants

Recently the use of data mining application in electricity power systems have been increased. Many papers were found in the literature, each is focusing in one area of the power system. Some are focusing in the distribution system like Ramos 2008, who used decision tree to classify the consumers. Saibal, 2008 [8] used WN (Wavelet Networking is an extension of perceptron networks) for the classification of transients. Figueiredo, 2005 [9] used decision tree for the classification of electric energy consumer. Dola, 2005 [10] used decision tree and neural network for faults classification in distribution system. Mori, 2002 [11] used regression tree and neural network for load forecasting. Other researcher focused in the transmission line problems like: Hagh 2007 [12] , Silva, 2006 [13] , Costa, 2006 [14] , Vasilic, 2002 [15] all of them used neural network to study faults detection, classification and locations in transmission lines. Dash, 2007 [16] used support vector machine for the classification and identification of series compensated. Vasilic 2005 [17] and Huisheng 1998 [18] used Fuzzy/ neural network for faults classification. Some other researchers focused in power generation part (power plants) like Andrew Kusiak et. al. [19] who used a multi objective optimization model to optimize wind turbine performance. Others like Ecir Ug˘ur Kucuksille et. al. [20] used data mining to predict thermodynamic properties. They used many algorithms to predict enthalpy, entropy and specific volume for specific types of refrigerants. Other researchers focused on work process optimization and performance monitoring. Water and Power Plant Fujairah (FWPP) in the United Arabic Emirates is a true success story of data mining usage, where more than 4% of of the total consumption have been achieved [21]. Softstat showed the superiority of data mining tools to traditional approaches like DOE (design of experiments), CFD (computational fluid dynamics). In their research they started by feature selection then apply data mining algorithms to get better performance of flame temperature, finally recommendations from the model was deployed [22].

OBJECTIVES

The first goal of this paper is to identify the parameters that influence the amount of power generated from thermal power plants, and their effect in the amount of generated power according to the current status of the power plant. The second goal is to build a prediction model that uses the selected parameters to predict the amount of power generated from the thermal power plant. The results of the proposed feature selection and prediction models could be used as a tool to diagnose the power plant problem

MATERIALS AND METHODS

In this paper the author followed CRISP-DM (Cross-Industry Standard Process for Data Mining) to build the prediction model. CRISP-DM is selected because it is a non-proprietary, documented, simple, well organized, and freely available data mining model, which was developed by industry leaders [23]. Waikato Environment for Knowledge Analysis (Weka) tool is used in this research because of its availability, simplicity, and it is suitable for similar researches.

As stated by CRISP-DM, we started by problem definition and data understanding, then data preparation was done. Three data sets were prepared one for each unit, and the third is a combined dataset to check whether a generic prediction model could be found. After that feature selection is done using wrapper method, to select the features that influence the amount of generated power. Then the selected features were used to build the prediction model. Because the outcome is numeric, regression algorithms were used to build the prediction model. Then obtained results were discussed, after that a conclusion is shown at the last part. Figure 2 shows the research framework, which shows the overall methodology followed in this research.

A) Data Collection and Pre-Processing

The datasets of this research were obtained from Khartoum North Thermal Power Plant KNTPP. This large (200 MW) power plant was commissioned in three phases, each phase is composed of two identical units, each unit is a separate power generation unit that follows Rankine Cycle . In this research we focus on Phase 2 which is composed of unit 3 and unit 4. Data is collected instantly by different types of sensors through SCADA system and recorded in a historical database. Due to disk space limitation, any data older than two months will be purged automatically from the database.

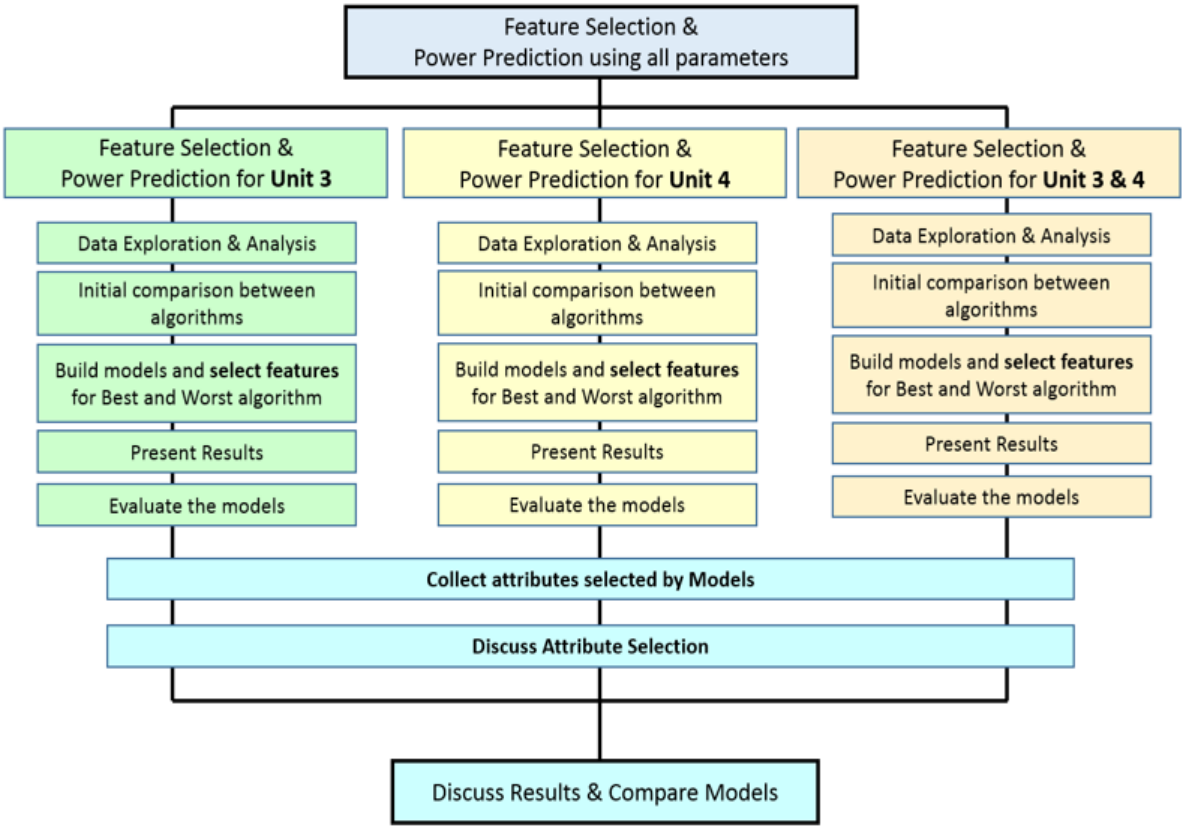

Figure 2 Research Framework

Efficiency engineers kept historical data for three years back (*between Aug-2012 and July-2015*), for plant efficiency analysis purposes. This monthly collected data is actually a snapshot of *2-minutes interval* readings for the first day of every month.

Data preprocessing is crucial to the integrity of data mining results. The raw data was reorganized in a relational format to be suitable for machine learning tools. Instances with null values were deleted from all datasets. The original datasets were composed of 83 features, these features were reduced to 63 according to domain expertise (*efficiency engineers*) decision. Table 1 shows the 63 features of the datasets.

B) The Datasets Characteristics

Three datasets were prepared for feature selection and building the power prediction models, one for each unit and the third is a generic dataset that combines unit 3 and unit 4 data in one set. The class in all datasets is: Generated power in Mega Watt(*GeneratedPowerM*).

All attributes including the class are numeric. Below is some description about these three datasets:

1. **Unit 3 dataset**: contains only 300 instances.
2. **Unit 4 dataset**: contains 720 instances.
3. **Unit 3&4 dataset**: This dataset is just a new file which combines both unit 3 and unit 4 datasets. So, the total number of incenses of this new dataset is 1020.

FEATURE SELECTION & PREDICTION MODEL

To achieve the goals of this research; three experiments were done, using three different datasets (*Unit 3* : to study unit 3 separately, *Unit 4*: to study unit 4 separately, *Unit 3&4* : is a combined dataset that contains both unit 3 and 4 data, to check whether a generic model could be used). As shown in figure 2 each experiment started by data exploration and analysis. Then an initial comparison is done between the seventeen regression algorithms provided by Weka, to select the algorithm that gives lowest *Mean Absolute Errors* and highest *Correlation Coefficient*

between attributes. The selected algorithm was then used to select the features that influence the amount of generated power, and build the regression model. More details about the experiment and its evaluation is shown in the rest of this part.

A) Data Exploration and Analysis:

Some statistical analysis is required to get more deep understanding about the datasets. Tables 2, 3 and 4 shows basic statistics about the most important attributes of Unit 3, Unit 4 and Unit 3&4 datasets respectively. The most important features are: Steam *Flow*, *pressure* and *temperature* of steam at turbine inlet and outlet, the class of all these datasets is the *GeneratedPowerinMW*. This basic data exploration gives an overview about the unit status. The standard deviation of *PressureInlet* is 15.059 in unit 3 compared to 0.909 in unit 4, this is caused by the maximum value of pressure at unit 3 which is 116.963 bar. This will make direct impact on the generated power values. Regarding *TemperatureOut*, in unit 3 it is very high compared to unit 4, the maximum value in unit 3 is 84.893, while its counterpart in unit 4 is 66.628. Even the mean value is higher, in unit 3 it is 67.473, while in unit 4 it is 51.65. Also there is big difference between minimum values of *TemperatureOut*, it is 47.99 in unit 3 and 40.434 in unit 4.

B) Initial comparison between Algorithms

Weka provides seventeen regression algorithms, all of them could be used for our datasets. To select the most suitable algorithm for each dataset an experiment had been designed. The experiment used the seventeen algorithms and applied them all for each dataset. Each experiment generated 5 evaluation factors: Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Root Relative Squared Error, and Correlation Coefficient. The best algorithm is the one that gives the lowest Mean Absolute Errors and the highest Correlation Coefficient between attributes. Table 5 shows experiments results for each dataset ordered by Correlation Coefficient descending, so the first row for each dataset is the most suitable algorithm for it.

C) Build the Models and Evaluate Results for Each Dataset

This section presents the details of feature selection process, and the prediction model creation. The following four steps were applied for each dataset:

i. Weka *AttributeSelectedClassifier* method was used for feature selection and building the prediction model in one step [1]. The algorithm that shows the highest correlation coefficient was selected as the classifier to build the prediction model. Wrapper method was used for feature selection, *ClasifierSubsetEval* algorithm was used as the evaluator, and the *Best First* method with *Bi-directional* option was used to search the total subsets. Test is done by splitting 66.0% of the dataset for training and the remaining for testing
ii. List of selected features for each dataset is presented.
iii. Evaluation results for each dataset is presented.
iv. A comparison graph that shows the actual and the predicted generated power is presented to visualize the model.

The subsequent sections shows the details of the above four steps, for each dataset.

1. Feature selection Result and Power Prediction Model for Unit 3 Dataset

According to the results of model evaluation experiments in table 5; the algorithm that shows the highest correlation co-efficient in Unit 3 dataset is *Pace Regression*. For feature selection *ClasifierSubsetEval* was used as the evaluator, and the Best First algorithm with Bi-directional option was used to search the total subsets. Figure 3 presented the selected model for Unit 3 in six blocks.

i. **Selected Features** : the first block at the top of figure 3 is the list of the selected features. Each feature has a unique ID which is shown in table 1. The selected features are marked by X letter under the ID feature and highlighted in green color. The number in the most top right cell (28), which is highlighted in yellow, means that: according to Pace Regression algorithm, only 28 features affect the amount of generated power.
ii. **Generated Power Prediction Model**: The block on the right side of figure 3 shows the model in a regression equation. The factor of each variable in the equation gives indication of how much this parameter has effect to determine the generated power.

iii. **Predicted vs Actual Graph**: the graph in the center of figure 3 shows the accuracy of the model, by plotting the actual and predicted values.
iv. **Model Evaluation**: the table at the lowest left side of figure (3) shows the results of evaluation of the test set which is done using Hold out/test set. The correlation coefficient between parameters is 0.9997, so the model accuracy is very high.
v. **Dataset info**: this block is just on the right side of the model evaluation block. It shows information about the dataset like: the number of training instances = 198, number of test instances=102, the algorithm which is Pace Regression, and the time required to build the model which is 5.89 seconds in our case.
vi. **Comments**: the last block is reserved area if any comments are required.

2. Feature selection Result and Power Prediction Model for for Unit 4 Dataset

The algorithm that shows the highest correlation co-efficient in Unit 4 dataset is *Linear Regression*. For feature selection *ClasifierSubsetEval* was used as the evaluator, and Best First algorithm with Bi-directional option was used to search the total subsets. Figure 4 also presents unit 4 model in the following six blocks:

i. **Selected Features** : 52 features were selected by this experiment, all of the most important 5 features (1,3,4,61,62) were selected.
ii. **Generated Power Prediction Model**: The factor of each variable in the equation gives indication about how much the variable affects the generated power. The value of pressure at turbine bleeders 1,2,3,4 and 5 is high, because these points are part of turbine outlet. Also the values of T_Aaxialdisplacement A, B and C are high, this is observation is *very interesting* as stated by domain experts. Such an observation could be communicated with maintenance team for more investigation.
iii. **Predicted vs Actual Graph**: it is very clear from the graph that predicted values are very accurate.
iv. **Model Evaluation**: model evaluation is done using Hold out/test set. The correlation coefficient is 0.9998, and the MAE is 0.0928. So the model accuracy is very high.
v. **Dataset info**: the number of training instances are 475, and number of test instances are 245, the algorithm which is Pace Regression, and the time required to build the model is 73 seconds.
vi. **Comments**: all important features were selected.

3. Feature selection Result and Power Prediction Model for Unit 3&4 dataset

The last dataset is the Unit 3&4, which is a combination of the two previous datasets. The purpose of this model is to check whether we can find a generic prediction model regardless of the unit status. The *Multilayer Perceptron* algorithm shows the highest correlation coefficient for unit 3&4 dataset. Although the correlation coefficient is 0.9997 and the MAE is 0.1294, but features 3 and 61 (which are belong to the most 5 important features group) were not selected by the model. So regardless the model accuracy it is not accepted. Figure 5 give all information about this model.

At least 5 models were created for each dataset, only the best model for each dataset was presented in this section. The features selection results of all models for each dataset is shown in table 6.

DISCUSSION

The domain experts stated that the most 5 important attributes are : (*1. Main steam flow, 3. Pressure at Turbine Inlet, 4. Temperature at Turbine Inlet, 61. Pressure at Turbine Outlet, 62. Temperature at Turbine Outlet.*), this group is called the *Top 5 group*. Elements of the Top 5 group are used to calculate the amount of generated power whether you are using manufacturer's consumption graph, or power equation. So, any model that failed to select these features will not be accepted. Table (7) shows results of the top 5 parameters, from this table the only algorithm that matches this constraint is Linear Regression for unit 4. Also this model attained the highest correlation coefficient (0.9998) and lowest MAE (0.0928) in unit 4, so Linear Regression model for unit 4 is the most acceptable model.

From Table (7) we can observe that feature 3 (Main steam header pressure) is not selected by any of the algorithms in unit 3 dataset, this observation directly related to the high StdDev

(15.059) which is observed for the same feature, see Table (3) Unit 3 Dataset Analysis. From this observation we can highlight that there is an issue in the amount of pressure at at turbine inlet of unit 3. It is the role of the domain expert to find interpretation for this observation.

Features Ranking

Tables 6 shows the ranking of features, the features that attained the highest ranking are : *1. Main steam flow* and *4. T/A inlet steam temperature*. Fortunately, both of these attributes belong to the Top 5 group, this is an evidence of the correctness of the feature selection results. Table 6 is considered to be the main analysis sheet which should be submitted to the domain experts to get more understanding and come up with the right diagnosis of the power plant problems.

Unit 3&4 Dataset

No algorithm in Unit 3&4 dataset succeeded to select the Top 5 features. Because this dataset is composed of both unit 3 and unit 4 dataset, it inherited all problems observed in unit 3. So, this dataset could not be used as generic dataset, and any unit should be studied separately to predict the generated power.

CONCLUSION

In reality things are different, although both units are identical at commissioning time, but each dataset shows different results, so we can neither depend on thermodynamic equations nor fabricant consumption graph to predict generated power "specially when power plant becomes older".

Data exploration and analysis is the initial and most important tool for power plant health check, that is very obvious from the high standard deviation found in Turbine Inlet Pressure of Unit 3 dataset.

Feature selection and prediction models can be used by efficiency engineers of power plant as a heath check tool to explore anomalies in the power plant, and to check how much specific attribute influence the power plant performance.

In order to come up with better results and proper innovations in such cases, it is better to form a research group from IT and electromechanical disciplines. Electromechanical engineers to define the problem and interpret the results, and IT engineers to prepare data and build the models.

ACKNOWLEDGEMENT

The author would like to thank all those who gave him the possibility to complete this article. Specially Prof. Izeldin Mohammed Osman whose help, stimulating suggestions and experience helped us during all the times of study. We are also grateful to Eng. Zuhier M. S. Daffaallah and Eng. Tarig from Khartoum North Power Plant who provided insight and expertise about power generation that greatly assisted the research.

REFERENCES

[1] Ian H. Witten, Frank Eibe, Mark A. Hall, (2011). *Data Mining : Practical Machine Learning Tools and Techniques 3rd ed*. Morgan Kaufmann New York

[2] Jefferson Morais, Yomara Pires, Claudomir Cardoso and Aldebaro Klautau (2009). *An Overview of Data Mining Techniques Applied to Power Systems*, Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.), ISBN: 978-3-902613-53-0

[3] htttp://www.saedsayad.com/data_mining_map.htm ((retrieved 3 Aug 2016)

[4] Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB conference, pp 487–499

[5] XindongWu ,Vipin Kumar , J. Ross Quinlan , Joydeep Ghosh , Qiang Yang , Hiroshi Motoda , Geoffrey J. McLachlan , Angus Ng , Bing Liu , Philip S. Yu , Zhi-Hua Zhou , Michael Steinbach ,David J. Hand , and Dan Steinberg (2008) .Top 10 algorithms in data mining . Knowl Inf Syst (2008) 14:1–37

[6] R K Kapooria, S Kumar, K S Kasana (2008). An analysis of a thermal power plant working on a Rankine cycle: a Theoretical Investigation. Journal of Energy in Southern Africa: 19:77-83

[7] www.learnengineering.org (retrieved 11 Nov 2015)

[8] Saibal Chatterjee, Sivaji Chakravorti, Chinmoy Kanti Roy and Debangshu Dey. (2008) Wavelet network-based classification of transients using dominant frequency signature, Electric Power Systems Research, Vol. 78, No. 1, (January 2008) 21-29.

[9] Figueredo, V.; Rodrigues F.; Vale, Z.; Gouveia, J. B. (2005). An eletric energy Consumer characterization Framework based on data mining techniques. IEEE Transactions Power Systems,

Vol. 20, No. 2., May 2005 , 596- 60), ISSN 1558-0679

[10] Dola, H.M.; Chowdhury, B.H. (2005). Data mining for distribution system fault classification. Power Symposium, 2005. Proceedings of the 37th Annual North American,pp. 457 – 462, ISBN 0-7803-9255-8, October 2005.

[11] Mori, H.; Kosemura, N.; Kondo, T.; Numa, K.; (2002). Data mining for short-term load forecasting. Power Engineering Society Winter Meeting. pp. 623 – 624, ISBN 0-7803-7322-7, Jan 2002

[12] Hagh, M.T.; Razi, K.; Taghizadeh, H. (2007). Fault classification and location of power transmission lines using artificial neural network, International Power Engineering Conference, pp. 1109 – 1114, ISBN 978-981-05-9423-7, Singapore, Dec 2007.

[13] Silva, K. M.; Souza, B. A.; Brito, N. S. D. (2006). Fault detection and classification in transmission lines based wavelet transform and ANN. IEEE Transaction on Power Delivery, Vol 21 , No. 4, (October 2006) 2058-2063, ISSN 0885-8977

[14] Costa, F. B.; Silva, K. M.; Souza, B. A.; Dantas, K. M. C.; Brito, N. S. D. (2006). A method for fault classification in transmission lines bases on ANN and wavelet coefficients energy. International Joint Conference Neural Networks. pp. 3700 – 3705, ISBN 0-7803- 9490-9, Vancouver, July-2006.

[15] Vasilic, S.; Kezunovic, M. (2002). An Improved Neural Network Algorithm for Classifying the Transmission Line Faults. IEEE Power Engineering Society Winter Meeting, pp.918 – 923. ISBN 0-7803-7322-7, Jan 2002.

[16] Dash, P.K.; Samantaray, S.R.; Panda, G. (2007). Fault Classification and Section Identification of an Advanced Series-Compensated Transmission Line Using Support Vector Machine. IEEE Transactions on Power Delivery, Vol 22. No. 22, (January 2007) 67 – 73, ISSN 0885-8977.

[17] Vasilic, S.; Kezunovic, M. (2005) Fuzzy ART Neural Network Algorithm for Classifying the Power System Faults. IEEE Transactions on Power Delivery, Vol. 20, No. 2, (April 2005) 1306-1314, ISSN 0885-8977.

[18] Huisheng Wang; Keerthipala, W.W.L. (1998). Fuzzy-neuro approach to fault classification for transmission line protection. IEEE Transactions Power Delivery, Vol.13, No. 4, 1093-1104, ISSN 0885-8977.

[19] Andrew Kusiak, Zijun Zhang, and Mingyang Li (2011). Optimization of Wind Turbine Performance With Data-Driven Models. IEEE Transactions On Sustainable Energy, 1:66-76.

[20] Ecir Ug˘ur Küçüksille, Res at Selbas , Arzu S encan (2011). Prediction of thermodynamic properties of refrigerants using data mining. ELSEVIER Energy Conversion and Management 52: 836–848.

[21] Himani Tyagi and Rajat Kumar (2014). Optimization of a Power Plant by Using Data Mining and its Techniques. International Journal of Advances in Science Engineering and Technology 2:83-87.

[22] www.statsoft.com (retrieved 29 Sep 2015)

[23] Colin Shearer (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data warehousing 16: 419 - 438.

Table 1: All Features of Power Prediction Datasets

#	Feature Name
1	***Main steam flow (kg/s)***
2	Total steam flow (kg/s)
3	***Main steam header pressure (bar)***
4	***T/A inlet steam temperature (deg C)***
5	Main steam header steam temperature (deg C)
6	HPH5 discharge feedwater flow (kg/s)
7	Feedwater temperature at economiser inlet (deg C)
8	Feedwater pressure at economiser inlet (bar)
9	Condenser right inlet temperature (deg C)
10	Condenser left inlet temperature (deg C)
11	Condenser right outlet temperature (deg C)
12	Condenser left outlet temperature (deg C)
13	Condensate water flow (kg/s)
14	Condenser hot well temperature (deg C) a
15	Condenser hot well temperature (deg C) b
16	Auxiliary steam flow (kg/s)
17	Auxiliary steam pressure (bar)
18	Auxiliary steam temperature (deg C)
19	Combustion air flow (Nm3/s)
20	Air temperature at FDF inlet (deg C)
21	Air temperature at FDF inlet (deg C)
22	Air temperature after RAH (deg C) (1)
23	Air temperature after RAH (deg C) (2)
24	FDF discharge air pressure (mbar)
25	FDF A speed (rpm)
26	FDF B speed (rpm)
27	Air temperature after SAH (deg C) (1)
28	Air temperature after SAH (deg C) (2)
29	HPH4 inlet feedwater temperature (deg C)
30	HPH4 outlet feedwater temperature (deg C)
31	HPH5 outlet feedwater temperature (deg C)
32	HPH5 inlet feedwater temperature (deg C)
33	T/A wheel champer steam pressure (bar)
34	T/A bleeder (1) pressure (bar)
35	T/A bleeder (2) pressure (bar)
36	T/A bleeder (3) pressure (bar)
37	T/A bleeder (4) pressure (bar)
38	T/A bleeder (5) pressure (bar)
39	T/A differential expansion (mm)
40	T/A axial displacement A (mm)
41	T/A axial displacement B (mm)
42	T/A axial displacement C (mm)
43	T/A bearing 3 vibration (mm/s)
44	T/A bearing 4 vibration (mm/s)
45	T/A bearing 1 vibration (1) (mic)
46	T/A bearing 1 vibration (2) (mic)
47	T/A bearing 2 vibration (1) (mic)
48	T/A bearing 2 vibration (2) (mic)
49	TBN side cold air (deg C)
50	TBN side warm air (deg C)
51	Exciter side cold air (deg C)
52	Exciter side warm air (deg C)
53	PMG side cold air (deg C)
54	PMG side warm air (deg C)
55	Generator winding temperature (deg C) 1
56	generator winding temperature (deg C) 2
57	Generator winding temperature (deg C) 3
58	Generator winding temperature (deg C) 4
59	Generator winding temperature (deg C) 5
60	Generator winding temperature (deg C) 6
61	***Condenser inlet exhaust steam pressure (bar a)***
62	***Condenser inlet exhaust steam temperature (deg C)***
63	**Generated power (MW)**

Table 2 Unit 3 Dataset Analysis

Statistic	SteamFlow	PressInlet	TempInlet	TempOut	PressOutlet	PowerMW
Minimum	46.879	34.686	498.064	47.99	0.085	39.482
Maximum	56.431	116.963	516.95	84.893	0.324	51.048
Mean	50.992	84.288	508.766	67.473	0.173	43.123
StdDev	2.429	15.059	2.83	10.715	0.068	3.746

Table 3 Unit 4 Dataset Analysis

Statistic	SteamFlow	PressInlet	TempInlet	TempOut	PressOutlet	PowerMW
Minimum	29.981	84.769	485.746	40.434	0.059	28.073
Maximum	60.601	95.13	524.338	66.628	0.225	57.358
Mean	46.06	87.866	508.304	51.65	0.113	43.498
StdDev	8.185	0.909	4.758	5.797	0.036	7.676

Table 4 Unit 3&4 Dataset Analysis

Statistic	SteamFlow	PressInlet	TempInlet	TempOut	PressOutlet	PowerMW
Minimum	29.981	34.686	485.746	40.434	0.059	28.073
Maximum	60.601	116.963	524.338	84.893	0.324	57.358
Mean	47.51	86.814	508.44	56.304	0.131	43.388
StdDev	7.353	8.354	4.286	10.461	0.055	6.762

Table 5 experiments results for for Unit 3, Unit 4 and Unit 3&4 datasets ordered by Correlation coefficient descending

Unit	#	Algorithm	Instances	Correlation coefficient	Mean absolute	Root mean squared	Relative absolute	Root relative squared	Time (s)
Unit3	2	PaceRegression	102	0.9997	0.0711	0.0949	2.23%	2.62%	5.89
Unit3	1	Linear Regression	102	0.9996	0.0798	0.1059	2.50%	2.93%	13.97
Unit3	3	SMOreg	102	0.9996	0.0723	0.1027	2.26%	2.84%	457.04
Unit3	4	MultilayerPerceptron	102	0.9995	0.0905	0.1182	2.84%	3.27%	373.43
Unit3	5	REPTree	102	0.9984	0.1606	0.2063	5.03%	5.70%	5.22
Unit4	1	Linear Regression	245	0.9998	0.0928	0.137	1.51%	1.81%	73.67
Unit4	2	Pace Regression	245	0.9998	0.1113	0.1529	1.81%	2.02%	28.23
Unit4	3	MultilayerPerceptron	245	0.9998	0.1216	0.164	1.98%	2.17%	1826.76
Unit4	4	DecisionTable	245	0.9917	0.3645	0.9718	5.94%	12.85%	121.72
Unit3 & 4	3	MultilayerPerceptron	347	0.9997	0.1294	0.1694	2.39%	2.50%	2894.61
Unit3 & 4	5	M5Rules	347	0.9996	0.1366	0.1811	2.52%	2.68%	752.64
Unit3 & 4	4	IBK	347	0.9981	0.2413	0.4147	4.46%	6.13%	43.99
Unit3 & 4	2	Linear Regression	347	0.9812	0.4281	1.3067	7.91%	19.31%	113.41
Unit3 & 4	1	Pace Regression	347	0.9808	0.4455	1.3194	8.23%	19.50%	22.97
Unit3 & 4	6	IsotonicRegression	347	0.9232	1.7774	2.6102	32.84%	38.58%	51.76

Table 6 Attribute selection results for Unit 3, Unit 4 and Unit 3&4

#		Unit 3					Unit 4				Unit 3 & 4						
	Algorithm >> / Feature	Linear Reg	PaceReg	SMO reg	M.L.Perceptron	REPTree	Linear Reg	Pace Reg	M.L.Perceptron	Decision Table	Pace Reg	Linear Reg.	M.L.Perceptron	IBK	M5Rules	Isotonic Reg	Total (Rank)
1	Main steam flow (kg/s)	1	1	1	1		1	1	1			1	1	1	1		11
4	T/A inlet steam temperature	1	1	1	1	1	1	1	1		1	1	1				11
37	T/A bleeder (4) pressure	1	1	1	1		1	1	1		1	1					9
2	Total steam flow (kg/s)	1	1	1	1	1	1	1				1					8
30	HPH4 outlet feedwater temperature	1					1	1			1	1	1		1	1	8
36	T/A bleeder (3) pressure	1	1	1			1	1			1	1	1				8
59	Generator winding temperature 5	1	1	1			1	1		1	1	1					8
62	Condenser inlet exhaust steam temp	1	1	1			1				1	1	1		1		8
13	Condensate water flow (kg/s)	1	1	1			1	1				1			1		7
14	Condenser hot well temperature a	1	1				1	1	1			1	1				7
16	Auxiliary steam flow (kg/s)		1	1			1	1	1			1	1				7
33	T/A wheel champer steam pressure	1	1	1	1		1	1				1					7
34	T/A bleeder (1) pressure	1					1	1		1	1	1			1		7
41	T/A axial displacement B (mm)		1	1	1		1	1			1	1					7
46	T/A bearing 1 vibration (2) (mic)	1	1	1	1		1					1			1		7
48	T/A bearing 2 vibration (2) (mic)	1	1	1	1		1	1				1					7
61	Condenser inlet exhaust steam press	1	1	1			1	1			1	1					7
18	Auxiliary steam temperature		1	1	1		1	1				1					6
22	Air temperature after RAH (1)	1	1	1			1	1				1					6
29	HPH4 inlet feedwater temperature				1	1		1			1	1	1				6
32	HPH5 inlet feedwater temperature	1					1	1		1	1	1					6
38	T/A bleeder (5) pressure	1				1	1	1	1			1					6
47	T/A bearing 2 vibration (1) (mic)	1	1	1			1	1				1					6
50	TBN side warm air						1	1	1		1	1			1		6
51	Exciter side cold air		1	1			1		1			1	1				6
52	Exciter side warm air	1	1	1	1		1					1					6
57	Generator winding temperature 3	1	1				1					1	1		1		6
5	Main steam header steam temp	1	1				1				1	1					5
9	Condenser right inlet temperarure	1				1	1					1	1				5
11	Condenser right outlet temperarure			1			1	1	1			1					5
12	Condenser left outlet temperarure	1					1	1	1			1					5
15	Condenser hot well temperature b	1		1			1	1				1					5
23	Air temperature after RAH (2)						1	1				1	1	1			5
28	Air temperature after SAH (2)	1		1			1	1	1								5
31	HPH5 outlet feedwater temperature	1		1			1				1	1					5
35	T/A bleeder (2) pressure					1	1	1				1	1				5
42	T/A axial displacement C (mm)	1					1	1				1	1				5
49	TBN side cold air	1	1		1		1					1					5
54	PMG side warm air	1					1		1			1			1		5
55	Generator winding temperature 1	1					1			1		1			1		5
56	generator winding temperature 2	1					1	1	1			1					5
58	Generator winding temperature 4	1					1	1	1			1					5
7	Feedwater temperature at economiser	1	1				1					1					4
8	Feedwater pressure at econom inlet				1	1	1					1					4
17	Auxiliary steam pressure			1			1	1				1					4
19	Combustion air flow (Nm3/s)	1	1					1			1						4
21	Air temperature at FDF inlet	1	1	1				1									4
25	FDF A speed (rpm)	1					1	1				1					4
26	FDF B speed (rpm)	1					1	1				1					4
40	T/A axial displacement A (mm)	1	1				1						1				4
43	T/A bearing 3 vibration (mm/s)					1		1	1			1					4
3	Main steam header pressure						1	1			1						3
10	Condenser left inlet temperarure	1		1								1					3
27	Air temperature after SAH (1)	1							1			1					3
39	T/A differential expansion (mm)	1				1	1										3
45	T/A bearing 1 vibration (1) (mic)						1					1	1				3
60	Generator winding temperature 6					1	1						1				3
6	HPH5 discharge feedwater flow	1											1				2
24	FDF discharge air pressure (mbar)	1										1					2
53	PMG side cold air	1					1										2
44	T/A bearing 4 vibration (mm/s)	1															1
Sum		47	28	26	13	10	52	38	16	4	16	51	18	2	10	1	

Table 7 Attribute selection summary of Unit 3 and 4 for the Top 5 features

		Unit 3					Unit 4				Unit 3&4						
#	Algorithm >> / Feature	Linear Reg.	PaceRegression	SMOreg	M.L.Perceptron	REPTree	Linear Regression	Pace Regression	M.L.Perceptron	Decision Table	Pace Regression	Linear Reg.	M.L.Perceptron	IBK	M5Rules	Isotonic Reg	Total (Rank)
1	Main steam flow (kg/s)	1	1	1	1		1	1	1			1	1	1	1		11
4	T/A inlet steam temperature	1	1	1	1	1	1	1	1		1	1	1				11
62	Condenser inlet exhaust steam temp	1	1	1			1				1	1	1		1		8
61	Condenser inlet exhaust steam press	1	1	1			1	1			1	1					7
3	Main steam header pressure						1	1			1						3

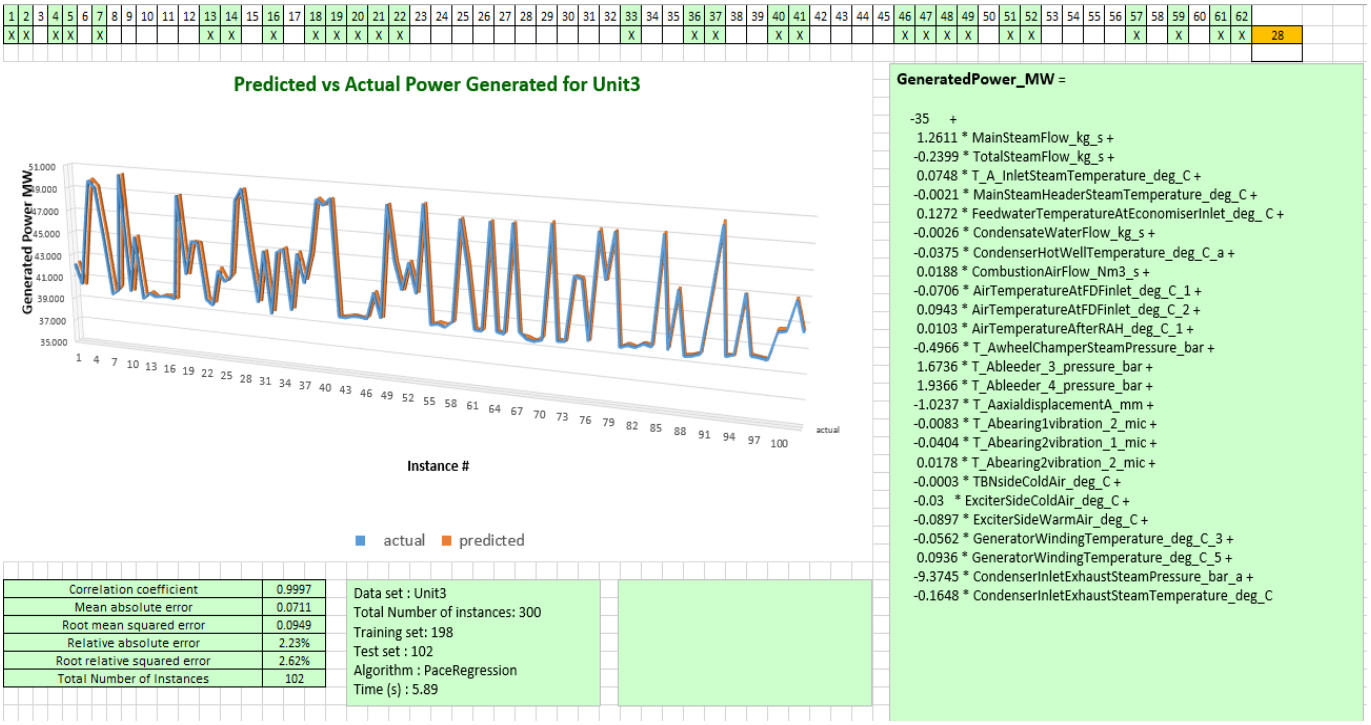

Figure 3 Generated Power model for Unit 3 using Pace Regression

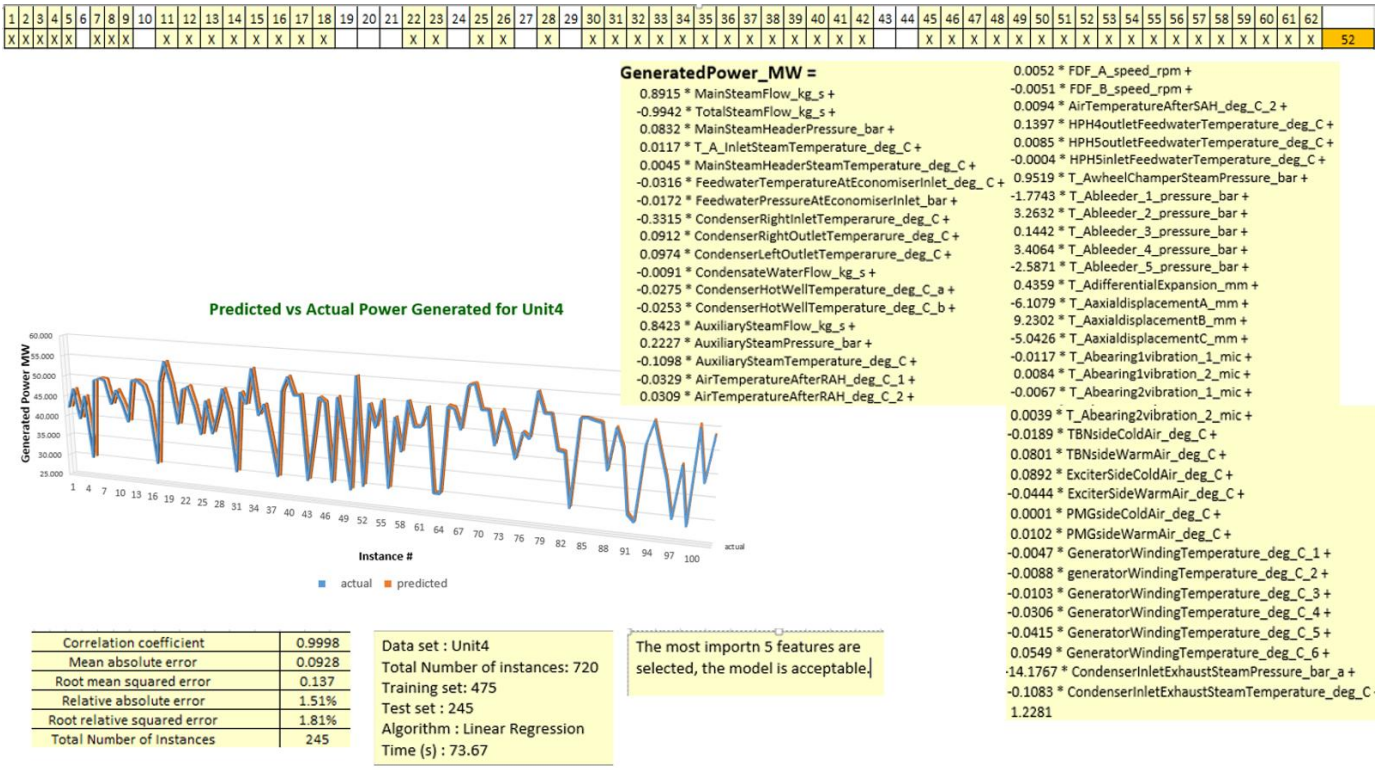

Figure 4 Generated Power model for Unit 4 using Linear Regression

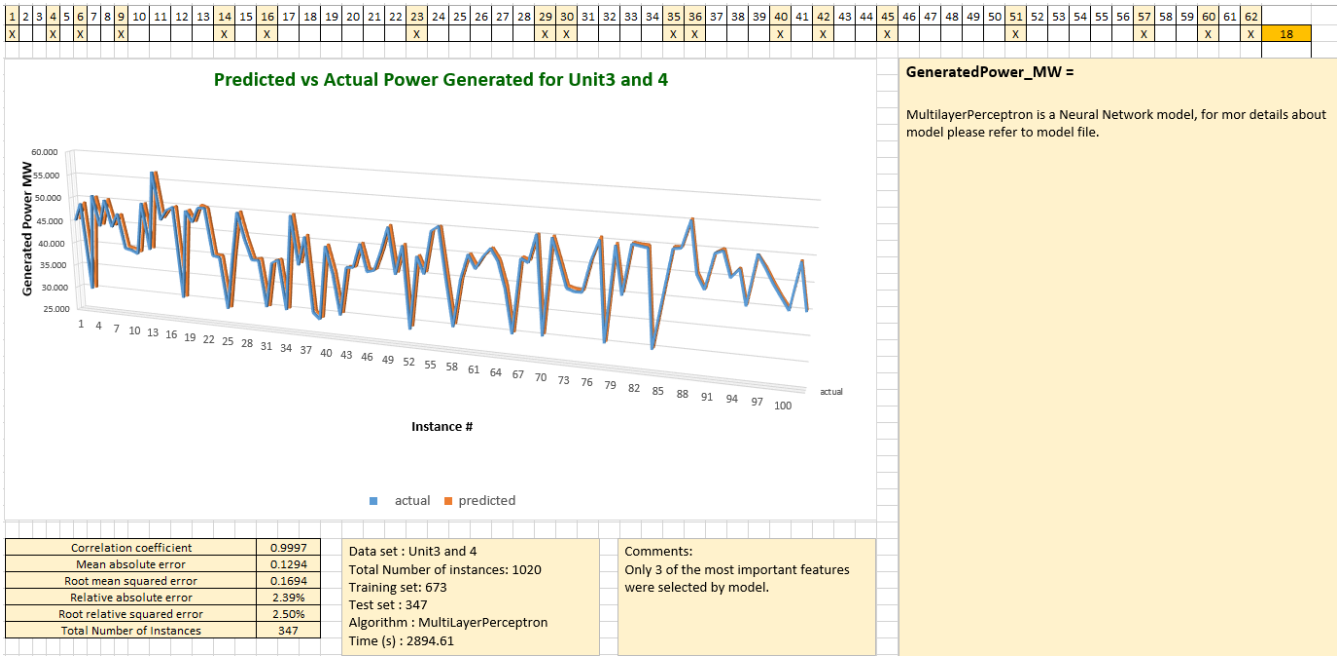

Figure 5 Generated Power model for Unit 4 using Multilayer Perceptron