Available online at www.uofk.edu/uofkej

UofKEJ Vol. 5 Issue 2 pp. 1-9 (August 2015)

UNIVERSITY OF KHARTOUM ENGINEERING JOURNAL (UOFKEJ)

Superiority of Data Mining Techniques to Predict the Amount of Power Generated by Thermal Power Plants

Waleed Hamed Ahmed Eisa, Naomie Bt Salim

Faculty of Computer Sciences, Sudan University of Science and Technology, Khartoum, Sudan (E-mail: mds.waleed@gmail.com)

Abstract: This paper presents the superiority of data mining techniques in predicting the amount of power generated by thermal power plants, over the traditional approaches that use thermodynamic laws or the power plant manufacturer's guides. The paper first compares between amount of power calculated using thermodynamic laws, and the amount of power predicted using manufacturers' guides with the actual power generated. Then prediction model was built to predict the amount of generated power using the controllable parameters at turbine inlet. Models were evaluated using separate test sets, or cross validation in case of small sets. The values predicted by this model is then compared with actual and other predicted values to prove that data mining tool is most accurate than traditional methods.

Keywords: *Power Plant; Thermal Power Plant; Power Prediction; Data Mining; Controllable Parameters.*

1. 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 is some obstacles that faces 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].

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 centres and the distribution which delivers power to the end users. There are many types of generators (also known as power plant) 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].

This research focus on thermal power plants that uses oil as energy source, these types of power plants uses Rankine Cycle to generate power [3]. Rankine Cycle is a closed system consists of four main components that are interconnected together to build one system, these components are shown in figure 1 are:

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

Recently the use of data mining application in electricity power systems has been increased. Some researchers focused in the distribution system, others focused in the transmission line, while other researchers studied power generation part (power plants) like Andrew Kusiak *et al.* [5] who studied the optimize of wind turbine performance. Others like Ecir Ug˘ur Kucuksille *et al.* [6] used data mining to predict thermodynamic properties. Other researchers focused on work process optimization and performance monitoring like [7]. Softstat [8] showed the superiority of data mining tools to traditional approaches like DOE (design of experiments), CFD (computational fluid dynamics).

There are two methods to predict the amount of generated power in thermal power plants. The first method is the expected output from the plant as stated by the manufacturer while the second method by using steam flow rate and enthalpy at turbine inlet and outlet [3]. Below are some details about each method.

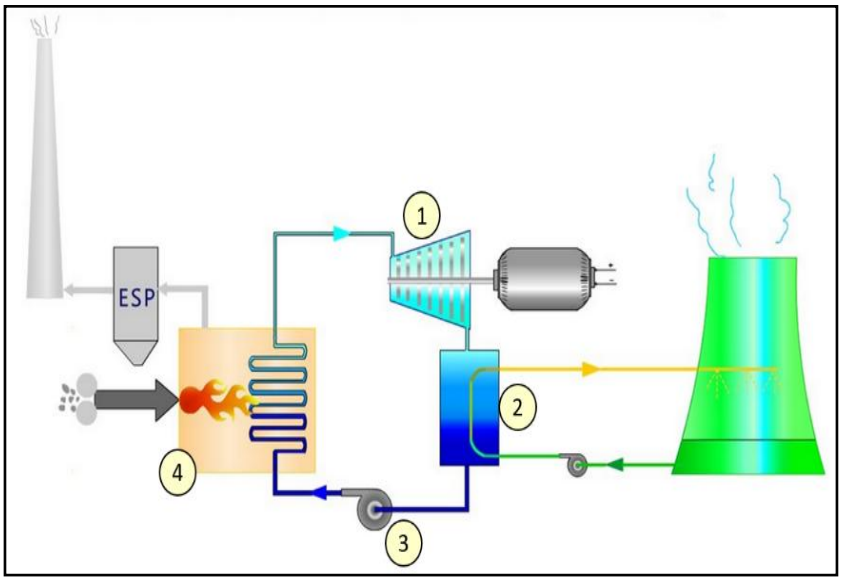

Fig. 1. Thermal power plant using Rankine Cycle [4]

A. Expected Amount of Generated Power according to manufacturer:

Upon power plant installation manufacturer provide the Steam Consumption Graph. It is a graph that shows how much power will be generated if steam with specific properties pumped to the turbine. **Fig. 2** shows the steam consumption graph for two units (Unit 3 and Unit 4) in Khartoum North Power Plant (a thermal power plant that is used as a case study in this research). From the steam consumption graph, a simple linear equation can be derived as the following:

$$Terminal\ Output[MW] = Live\ Steam\ Flow[kg/s] - 2 \quad (1)$$

B. Basic Equation of power calculation

It is known that the amount of power generated from thermal power plant could be calculated as follows [3]:

$$Power\ [MW] = ms \times (\dot{h_{in}} - h_{out}) \quad (2)$$

where : $\dot{m}s$: is the flow rate of steam at turbine inlet.
h_{in} : is enthalpy at turbine inlet.
h_{out} : is enthalpy at turbine outlet

In this paper the author followed CRISP-DM to build the prediction model [9]. Despite the two methods used to calculate the amount of generated power are seem to be simple and direct, however both of them become inaccurate when the plant becomes older. The first method is defined by manufacturer at commissioning time, but when plant becomes older normally things went different, because the performance of many parts degraded. The second method depends on a theoretical equation and should be multiplied by the efficiency factor of the power plant which is subject to the applied cycle and many other factors. Because of all these reasons the actual generated power is normally different from the two values.

Sample of the difference between these three values (amount expected according to fabricants, amount calculated using power equation and actual amount) for two different units (Unit 3 and Unit 4 in KNPP) is shown in Figs 3 and 4 respectively.

The goal of this research is to use data mining tools to find more accurate way to predict the amount of generated power in thermal power plants, and to prove that each unit in the power plant is totally different from others and no generic method could be used to predict power in all units, specially when the power plant becomes older.

2. MATERIALS AND METHODS

In this part; first the methodology followed will be shown, then a brief description of the datasets will be presented.

2.1. Methodology

There are hundreds of parameters that are captured at different points of the power plant, very few of them are controllable. In this research we focused only on three controllable parameters that directly influence the amount of generated power. These controllable parameters are :

- Steam Flow Rate at turbine inlet.
- Steam Pressure at turbine inlet.
- Steam Temperature at turbine inlet.

So the goal of this research is to build a prediction model to predict the amount of generated power using these three parameters. To build the model, two datasets (Unit 3 and Unit 4) from two different units in Khartoum North Power Plant were used. **Fig. 3** shows the general framework of this research. Below is the exact steps that were followed to build and evaluate the two models:

- Weka Explorer was used to specify the algorithms that could be sued with the available datasets. The data types

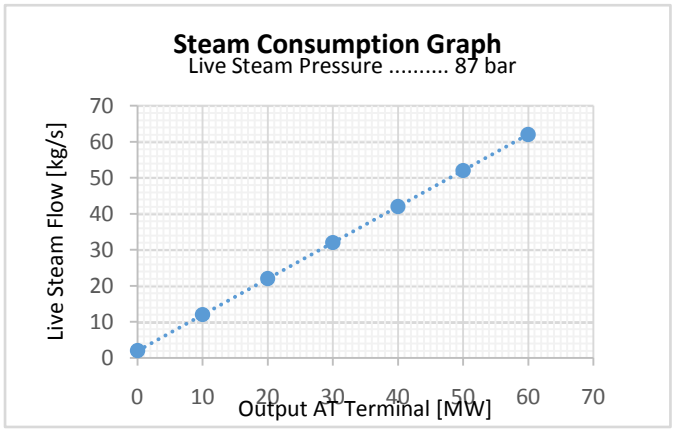

Fig. 2. Steam Consumption Graph for KNPP

of predictors and classes of all datasets are numeric. So, only regression algorithms will be valid for these datasets. Although more than 20 algorithms are available in Weka for regression, only 17 are usable. After a preliminary test some algorithms were excluded because some of them give different errors, while the others took very long time to build the model. Table (1) shows the valid algorithms for our datasets

- Then an initial comparison between algorithms was done using Weka Experimenter
- Then prediction models were built using the selected algorithm from the previous step, and obtained results were evaluated.
- Finally results were discussed.

2.2 Datasets Description

2.2.1 Data Collection

The datasets used in this research were obtained from Khartoum North Thermal Power Plant KNPP. 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.

Raw data is collected instantly by different types of sensors through SCADA system and recorded in a historical database. Due to disk space limitation, data older than two months automatically purged from the database. For efficiency analysis purposes one hour snapshot (*of 2-min interval*) is taken every month during the last three years (*between Aug-2012 and July-2015*). All these snapshots were taken from the two units.

2.2.2 *Data Pre-Processing*

Data preprocessing is crucial to the integrity of data mining results. To prepare the datasets for this research, the following steps were done:

- *Select controllable parameters*: Three controllable parameters (Steam Flow, Temperature at Turbine Inlet, pressure at Turbine Inlet) were selected to be the *predictors* and the actual generated power is the *target*.
- *Select sample instances*: normally, changes in data collected at the same day is very little, so only three instances were selected from each day to build the dataset for each unit.
- *Calculate generated power for comparison*: two values of generated power were calculated using two different methods:

  a) *Expected amount according to Fabricant*: this value is calculated using the steam consumption graph (figure (2)) which is provided by fabricants.
  b) *Expected amount as per thermodynamic Equations:* is calculated using the main steam flow rate at turbine inlet and enthalpy at turbine inlet and outlet.

The amount of generated power calculated using above methods will be compared later with : the *actual* amount of the generated power, and the amount *predicted* using the selected data mining model.

.2.1 Datasets

Two datasets were prepared to build the power prediction models, one for each unit, table (2) and (3) shows a sample of Unit 3 and unit 4 datasets respectively. Each dataset contains 87 instances. The class in all datasets is the **Actual (** *The Actual Generated Power in Mega Watts***)** . All attributes including the class are numeric.

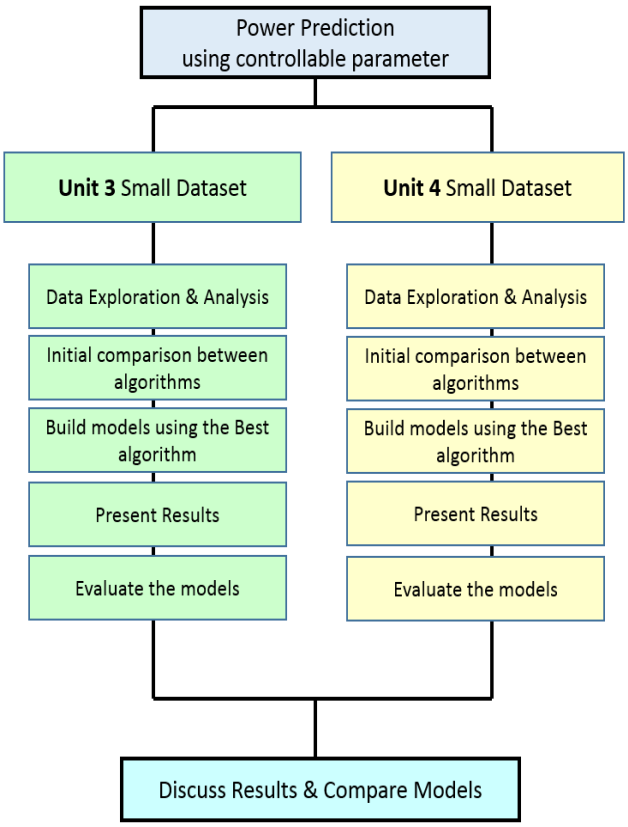

Fig. 3. Research Framework

Table 1. Valid Regression Algorithms for our datasets

Classifier	#	Algorithm
Functions	1	GaussianProcess
	2	IsotonicRegression
	3	LastMedSq
	4	LinearRegression
	5	MultilayerPerceptron
	6	PaceRegression
	7	SimpleLinearRegression
	8	SMOreg
Lazy	9	IBK
	10	Kstar
	11	LWL
Rules	12	ConjunctiveRule
	13	DecisionTable
	14	M5Rules
Trees	15	DecisionStump
	16	M5P
	17	REPTree

Table 2. Sample of Unit 3 dataset

Main Steam Flow	Pressure Inlet	Temperature Inlet	Actual
27.266	86.058	507.502	26.022
29.456	85.823	507.956	28.073
30.215	86.493	505.154	28.327
29.794	85.84	506.907	28.464

Table 3. Sample of Unit 4 dataset

Main Steam Flow	Pressure Inlet	Temperature Inlet	Actual
27.266	86.058	507.502	26.022
29.456	85.823	507.956	28.073
30.215	86.493	505.154	28.327
29.794	85.84	506.907	28.464

Table 4. Unit 3 Dataset Analysis

Statistic	Steam Flow	Press Inlet	Temp Inlet
Minimum	24.569	0.4	498.064
Maximum	59.373	128.782	530.501
Mean	45.402	75.291	508.93
StdDev	8.153	28.875	4.203

Table 5. Unit 4 Dataset Analysis

Statistic	Steam Flow	Press Inlet	Temp Inlet
Minimum	27.266	84.804	493.465
Maximum	60.021	91.042	524.338
Mean	45.064	87.675	509.902
StdDev	9.322	1.198	4.8

2.3 *Datasets*

Two datasets were prepared to build the power prediction models, one for each unit, table (2) and (3) shows a sample of Unit 3 and unit 4 datasets respectively. Each dataset contains 87 instances. The class in all datasets is the **Actual (***The Actual Generated Power in Mega Watts***)**. All attributes including the class are numeric.

4. PREDICTION MODELS

To build the prediction models, the steps of research frame work which are shown in figure (3) were followed.

4.1 Data Exploration and Analysis

Some statistical analysis is required to get deep understanding about the datasets. Tables 4, and 5 show basic statistics about attributes of unit 3 and unit 4 datasets respectively. This basic analysis gives basic ideas about the status of these units. It is clear that Std Dev of *Pressure Inlet* of unit 3 is high, it is 28.875 compared to 1.198 in unit 4, this is caused by the maximum value of pressure at unit 3 which is 128.782 bar. This will make direct impact on the amount of the generated power value. The mean value of pressure in unit 3 is 75.291 bar, while the optimum value for pressure (as specified by fabricants) should be 87 bar. Unit 4 statistics is normal and very near to proposed values by fabricants.

4.2 Initial comparison between Algorithms

The purpose of this step is to make initial comparison between the 17 valid algorithms to select the best one for each dataset. Because we have two different datasets; two experiments were done "one for each dataset". Each experiment generate 5 evaluation factors: *Mean absolute error*, *Root mean squared error*, *Relative absolute error*, *Root relative squared error* and *Correlation coefficient*. Table 6 and 7 show the initial comparison results for unit 3 and 4 respectively. Correlation coefficient is used to rank values in the two tables in descending order, so the highest row in each unit is the best performance model.

4.3 Build the Models and Evaluate Results for Each Dataset

In this section the details of the models' creation will be presented. For each dataset two models were built using the best and worst algorithms:

The best algorithm model: The algorithm that shows the highest correlation coefficient was used to build the model using the following steps:

1. Weka Explorer was used with the best algorithm to build the model.
2. Then evaluation results for the model were presented.
3. Finally, a graph was plotted to compare between the actual and the predicted amount of generated power to reflect the model accuracy.

The worst algorithm model: The algorithm that shows the lowest correlation coefficient was also used to build another model for comparison, using the same above steps.

Details for each model will be presented in a separate section, then comparison between models and discussion about results will be the last section of this part.

4.3.1 Power Prediction Model using controllable parameters of unit 3 dataset

According to the results of model evaluation experiments in Table 6; the algorithm that shows the highest correlation co-efficient in Unit 3 dataset is Pace Regression. Weka Explorer is used with Pace Regression to create the model. Because the dataset is small, evaluation is done using 10-fold cross validation. Complete information about the model is presented in **Fig. 6**, which is composed of 5 blocks:

- *Generated Power Prediction Model*: The block on the right side of the figure shows the model in a form of equation that uses the three controllable parameters.
- *Predicted vs Actual Graph*: the graph in the center of the figure shows the accuracy of the model, by plotting the actual and predicted values. It is clear from the graph, that the two lines are not identical.
- *Model Evaluation*: the table at the lowest left side of the figure shows the results of evaluation using 10-fold cross validation. The correlation coefficient is 0.9383, which means that the model is not so accurate.
- *Dataset info*: this block is on the right side of the Model evaluation block. It shows information about the dataset, number of training instances is 87, evaluation method is 10-fold cross validation, the algorithm is Pace Regression, and the time required to build the model which is 0 seconds in our case.
- *Comments:* the last block is reserved area if any comments is required.

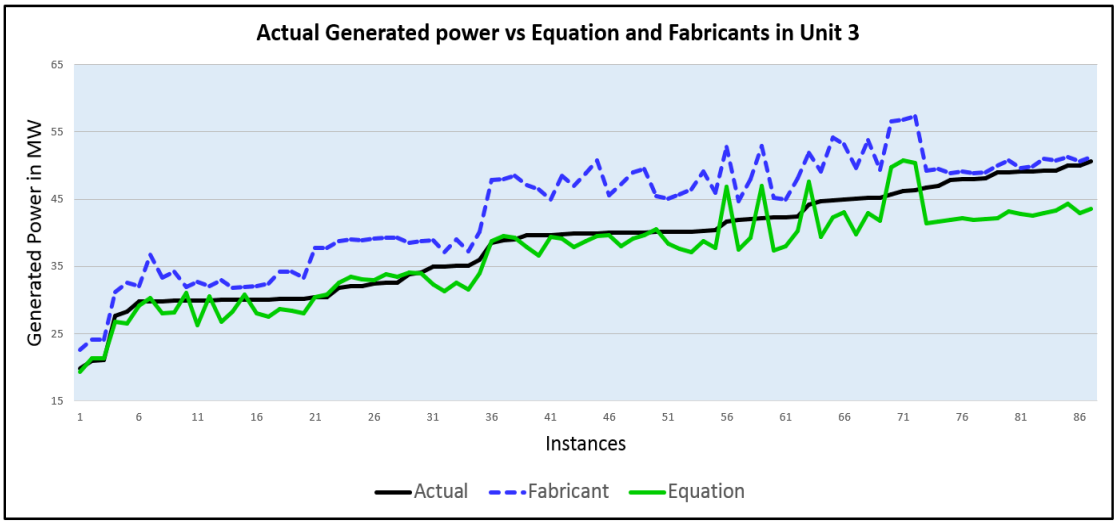

Fig. 4. Actual Generated Power versus Equation and Fabricants in Unit 3

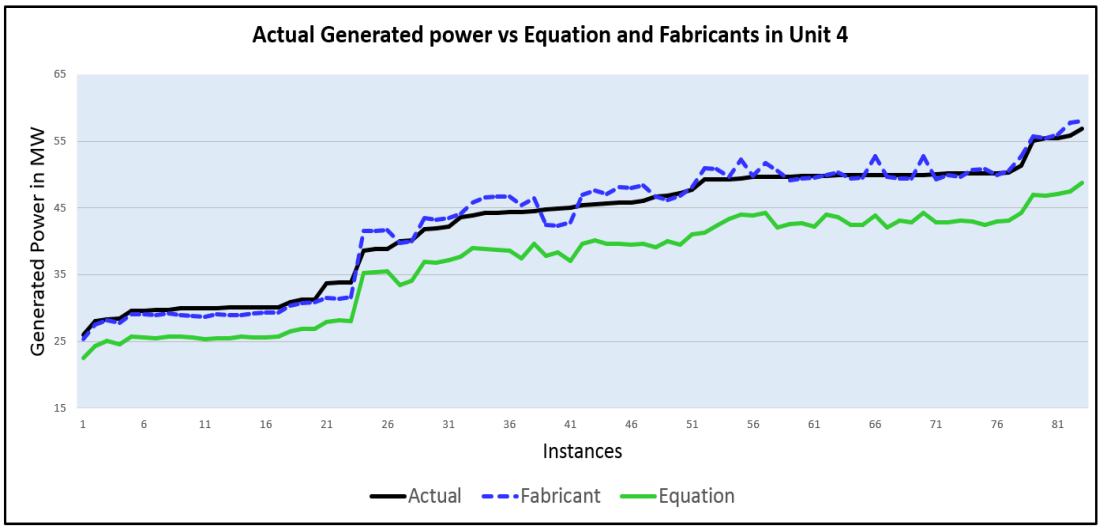

Fig. 5. Actual Generated Power versus Equation and Fabricants in Unit 4

Table 6. Experiment Results to compare Regression Algorithms for Unit 3 dataset

No.	Algorithm	Correlation coefficient R	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
6	weka.classifiers.functions.PaceRegression	0.9383	2.1045	2.598	32.5636	34.2217
1	weka.classifiers.functions.GaussianProcesses	0.9381	2.2890	2.8501	35.4179	37.5422
16	weka.classifiers.trees.M5P	0.9379	2.0848	2.6059	32.2593	34.3258
14	weka.classifiers.rules.M5Rules	0.9378	2.0973	2.6106	32.4528	34.3874
4	weka.classifiers.functions.LinearRegression	0.9375	2.1031	2.6128	32.5425	34.4169
8	weka.classifiers.functions.SMOreg	0.9374	2.1269	2.6592	32.9098	35.0282
2	weka.classifiers.functions.IsotonicRegression	0.9371	1.9029	2.6219	29.4443	34.5362
7	weka.classifiers.functions.SimpleLinearRegression	0.9241	2.3287	2.8694	36.0325	37.7969
13	weka.classifiers.rules.DecisionTable	0.9216	2.2394	2.9157	34.6513	38.4061
9	weka.classifiers.lazy.IBk	0.9189	1.6682	2.9801	25.8128	39.2545
17	weka.classifiers.trees.REPTree	0.9164	2.1505	3.0152	33.2754	39.7166
3	weka.classifiers.functions.LeastMedSq	0.9015	2.6101	3.3594	40.3865	44.2514
5	weka.classifiers.functions.MultilayerPerceptron	0.8998	2.6154	3.3355	40.4688	43.9355
10	weka.classifiers.lazy.KStar	0.8856	2.4211	3.7060	37.4625	48.8163
11	weka.classifiers.lazy.LWL	0.8652	3.0524	3.7723	47.2315	49.6903
12	weka.classifiers.rules.ConjunctiveRule	0.8636	3.0787	3.7870	47.6377	49.8832
15	weka.classifiers.trees.DecisionStump	0.8558	3.1496	3.8868	48.7345	51.1977

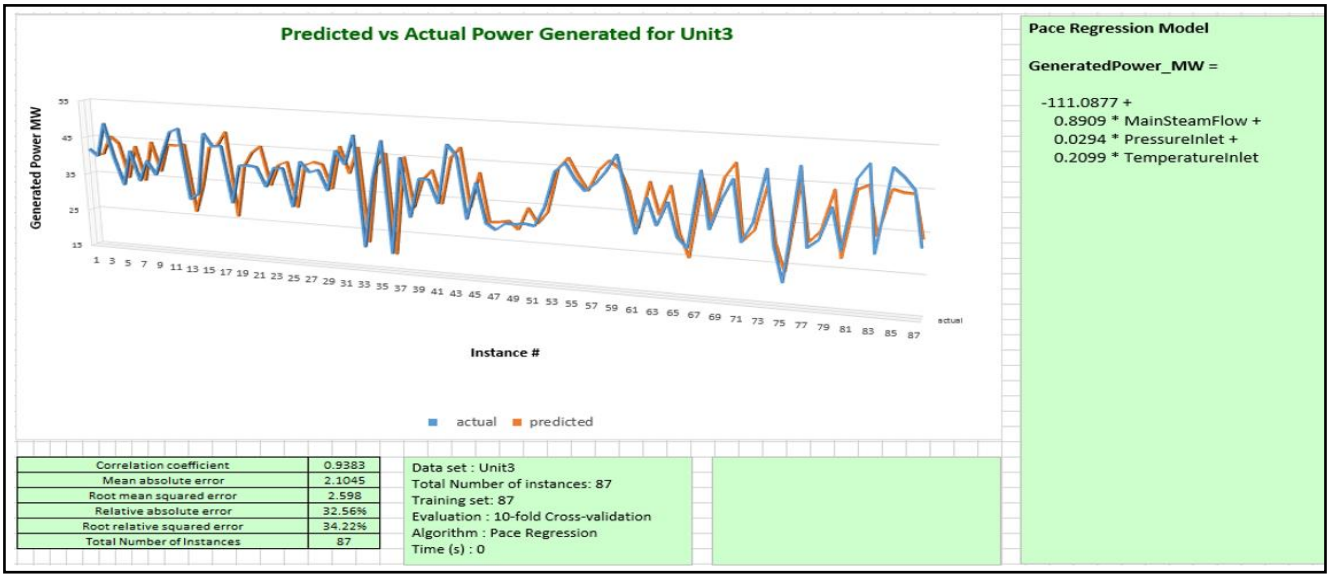

Fig. 6. Generated Power model for Unit 3 using Pace Regression

4.3.2 Power Prediction Model using controllable parameters dataset for Unit 4

According to the results of model evaluation experiments in Table 7; the algorithm that shows the highest correlation co-efficient in Unit 4 dataset is Isotonic Regression. Weka Explorer is used with Isotonic Regression to create the prediction model using the controllable parameters. Model evaluation is done using 10-fold cross validation. Complete information about the model is presented in Fig. 7, which is composed of 5 blocks:

- *Generated Power Prediction Model*: the model shown in **Fig. 7** based on steam flow attribute, it was built using Isotonic Regression mode, so the interpretation of the model is not clear to normal users because it is not written as a normal formula.
- *Predicted vs Actual Graph*: it is very clear from the graph that predicted values are accurate, and much better than those of unit 3. Still the model accuracy is less than what was obtained when full feature dataset

was used, that is because this model is more realistic and is using only the controllable parameters.

– *Model Evaluation*: Model evaluation is done using 10-fold cross validation. The correlation coefficient is 0.9943, and the MAE 0.644 so the model accuracy good.
– *Dataset info*: the number of training instances = 83, the algorithm which is Isotonic Regression, and the time required to build the model is 0.01 seconds.
– *Comments*: model is accurate.

Table 7. Experiment Results to compare Regression Algorithms for Unit 4 dataset

No.	Algorithm	Correlation coefficient, R	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
2	weka.classifiers.functions.IsotonicRegression	0.9943	0.6440	0.9224	8.5690	10.6071
8	weka.classifiers.functions.SMOreg	0.9915	0.8501	1.1267	11.3126	12.9567
3	weka.classifiers.functions.LeastMedSq	0.9910	0.9107	1.1893	12.1188	13.6770
4	weka.classifiers.functions.LinearRegression	0.9910	0.9157	1.1534	12.1845	13.2638
6	weka.classifiers.functions.PaceRegression	0.9910	0.9073	1.1562	12.0728	13.2964
7	weka.classifiers.functions.SimpleLinearRegression	0.9896	1.0013	1.2377	13.3236	14.2333
14	weka.classifiers.rules.M5Rules	0.9894	0.9533	1.2522	12.6861	14.4000
16	weka.classifiers.trees.M5P	0.9893	0.9648	1.2571	12.8379	14.4567
17	weka.classifiers.trees.REPTree	0.9893	0.8271	1.2592	11.0057	14.4813
13	weka.classifiers.rules.DecisionTable	0.9881	0.9644	1.3264	12.8326	15.2538
10	weka.classifiers.lazy.KStar	0.9864	0.9616	1.4347	12.7954	16.4986
5	weka.classifiers.functions.MultilayerPerceptron	0.9857	1.1365	1.4871	15.1233	17.1016
1	weka.classifiers.functions.GaussianProcesses	0.9854	1.4071	1.6934	18.7246	19.4742
9	weka.classifiers.lazy.IBk	0.9810	1.0184	1.6711	13.5519	19.2180
11	weka.classifiers.lazy.LWL	0.9210	2.5788	3.3608	34.3166	38.6498
15	weka.classifiers.trees.DecisionStump	0.9006	3.0655	3.7443	40.7924	43.0597
12	weka.classifiers.rules.ConjunctiveRule	0.8902	3.1135	3.9343	41.4317	45.2443

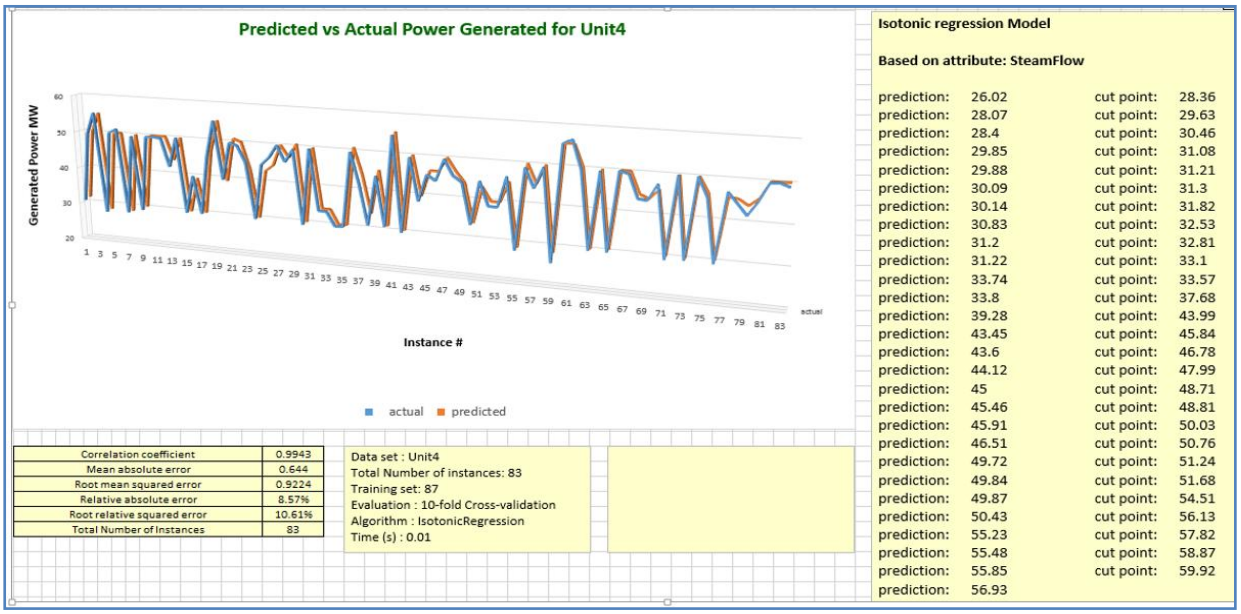

Fig. 7. Generated Power model for Unit 4 using Isotonic Regression

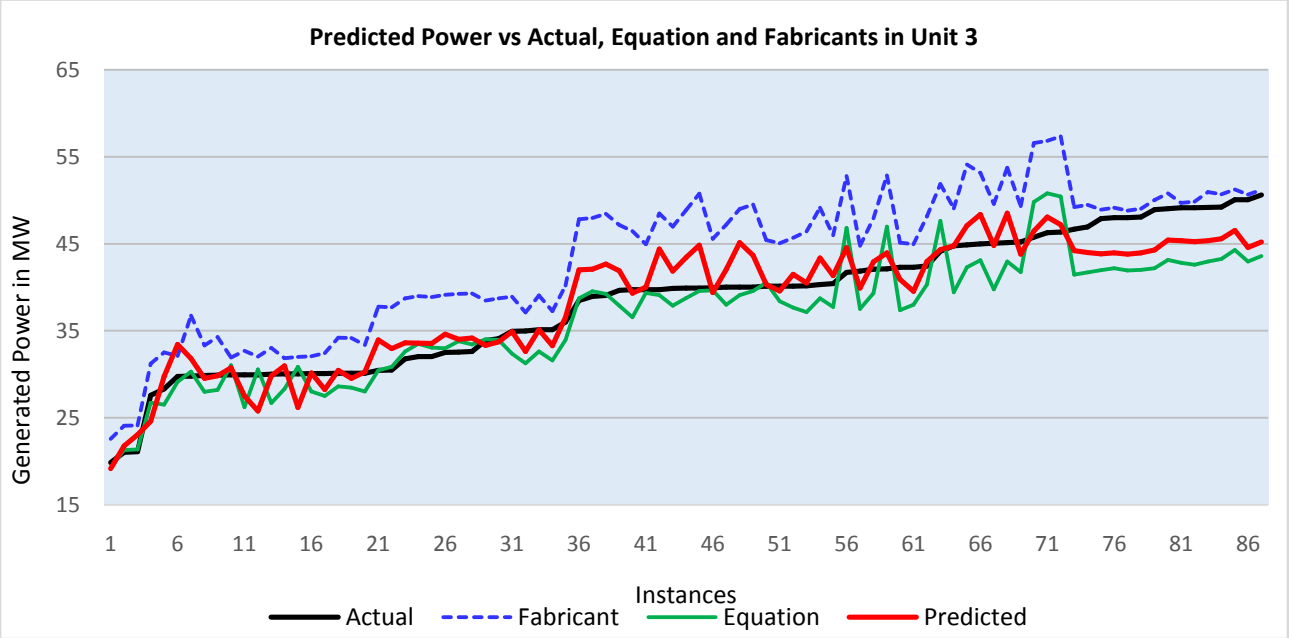

Fig. 8. Predicted power versus (Actual, Equation and Fabricants) for Unit 3

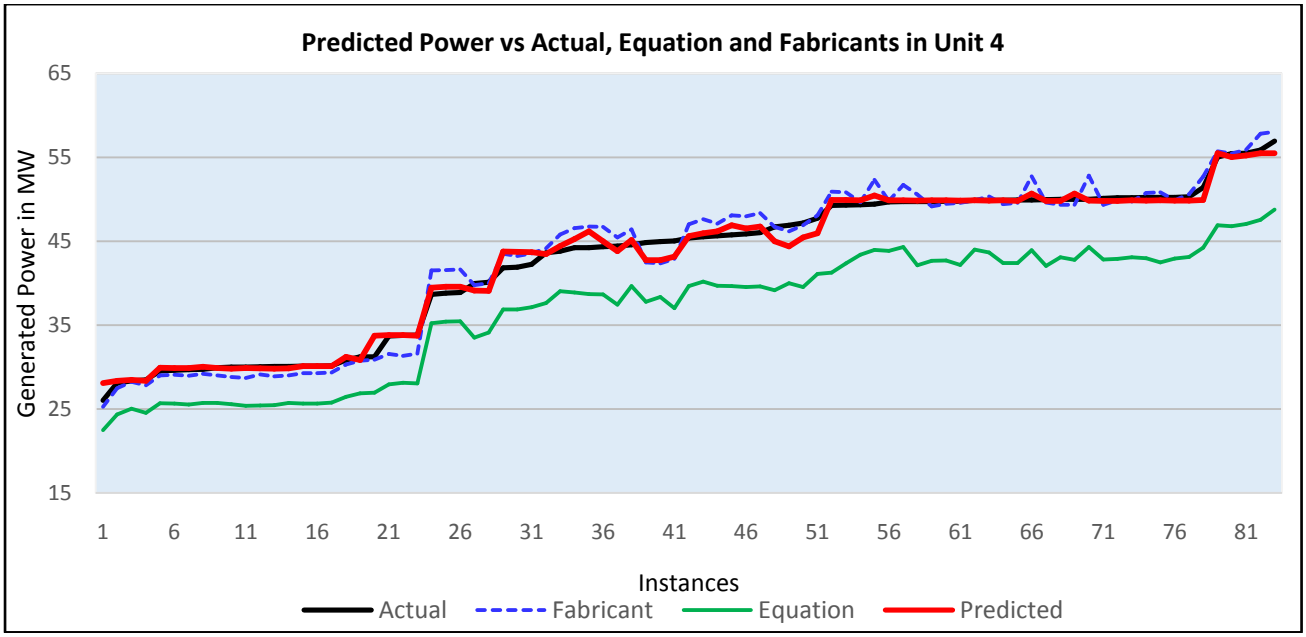

Fig. 9. Predicted power versus (Actual, Equation and Fabricants) for Unit 4

5. RESULTS DISCUSSION AND MODELS COMPARISON

Unit 3 Results

As discussed in Data Exploration and Analysis above In unit 3, a very high and a very low pressure observed (128 and 0.4 bar), this big difference caused a high value of standard deviation of pressure (28.875). Because pressure at turbine inlet is one of the most important parameters that are used to predict the amount of generated power, this difference leads to high difference in the **actual** amount of generated power. Also pressure is one of the model **predictors**, subsequently the correlation coefficient is low (0.9383) and the Mean Absolute Error is high (2.1045).

Although there is a clear problem in one of the predictors, the value predicted by Pace Regression model is much better than the value predicted by equation and expected by fabricants. Fig. 8 shows a comparison between amount of power predicted by Pace Regression model versus (Actual, Equation and Fabricants values).

Unit 4 Results

Unit 4 dataset is much better than unit 3, that is obvious from the table(4) which shows basic statistical analysis of unit 4 dataset. The mean values of pressure and temperature is very near to optimum values assigned by fabricants. So, the amount of the actual generated power is very near to the amount expected by fabricants. Figure (9) shows a comparison between amount of power predicted by Isotonic Regression model versus (Actual, Equation and Fabricants values). It is also very clear that the predicted value is much accurate than the values calculated by equation or expected by fabricants.

Did the model succeeded to answer the research questions?

Yes, from the above discussion of model evaluation results, it is clear that; values predicted by both unit 3 and 4 is defiantly better than those calculated by equation or expected by fabricants. Moreover, the new models depends only on controllable parameters. These models (Pace Regression for unit 3 and Isotonic Regression for unit 4) could be used to predict the amount of generated power accurately.

An additional contribution of this research is that : by data exploration and analysis and prediction model; the root cause of unit 3 problem had been highlighted.

6. CONCLUSIONS

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

The methodology used by this research is generic, subsequently it could be applied to any steam power plants to predict the amount of generated power accurately.

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.

In order to come up with better results and proper innovations in such a research, 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.

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] 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

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

[5] 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.

[6] 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.

[7] 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.

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

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