Predicting the power of a combined cycle power plant using machine learning methods

Conference Paper · November 2020

DOI: 10.1109/CCCI49893.2020.9256742

5 authors, including:

Salama Alketbi
University of Sharjah
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Ismail Shahin
University of Sharjah
85 PUBLICATIONS **724** CITATIONS
SEE PROFILE

Ali Bou Nassif
University of Sharjah
115 PUBLICATIONS **1,506** CITATIONS
SEE PROFILE

Ashraf M Elnagar
University of Sharjah
99 PUBLICATIONS **795** CITATIONS
SEE PROFILE

Predicting the power of a combined cycle power plant using machine learning methods

Salama Alketbi
Department of Computer Engineering
University of Sharjah
Sharjah, UAE
u15100348@sharjah.ac.ae

Ali Bou Nassif
Department of Computer Engineering
University of Sharjah
Sharjah, UAE
anassif@sharjah.ac.ae

Maha Alaa Eddin
Department of Computer Engineering
University of Sharjah
Sharjah, UAE
malaaeddin@sharjah.ac.ae

Ismail Shahin
Department of Electrical Engineering
University of Sharjah
Sharjah, UAE
ismail@sharjah.ac.ae

Ashraf Elnagar
Department of Computer Science
University of Sharjah
Sharjah, UAE
ashraf@sharjah.ac.ae

***Abstract*—The gas turbine is the most important part of the combined cycle power plant that generates the total electric power from the fuel to provide it to the houses, schools, and other facilities in the country. Thus, it is important to predict the power to increase and maximize profit. This paper compares four machine learning algorithms which are Multiple linear Regression, Multilayer perceptron, K- Nearest Neighbors, and Random Forest Algorithm. The dataset consists of 9,568 observations and four inputs which are ambient temperature, ambient pressure, relative humidity, and exhaust vacuum that will be used to train the prediction of the total electric power consumption which is the output. The best result was shown by using the Random Forest Algorithm with the mean absolute error of 2.3013 and root mean square error with 3.3061.**

Keywords—K-Nearest neighbors, Power prediction, Multilayer Perceptron, Multiple linear regression, Random forest

I. INTRODUCTION AND MOTIVATION

Generating electricity extracted from the fuel goes through several phases. This can be achieved in a combined cycle power plant. This type of technology will produce two types of energy electricity and steam. Combining the cycles will generate 50% more than the single cycle technology [1]. First, the gas will burn in Gas Turbine and mixed with the air that comes from the air filter. That combination will spin the generator and by its turn will generate the electricity. The heat lost from the gas turbine will be captured in the Heat Recovery Steam Generator (HRSG). An HRSG creates steam using boiled water that will spine an additional turbine generator to produce electricity. Finally, the stem will become liquid to recital it by using a condenser [2]. Figure 1 shows the combined cycle power plant processes.

Many countries establish this technology to generate more electricity due to the high demand to consider the environment and economic aspects. The main part of this plant is a gas turbine that will generate power.

The data set presented consists of four inputs temperature (AT), ambient pressure (AP), relative humidity (RH) and exhaust vacuum (V) that will be used to predict the total electric power (EP) hourly which is the output. The dataset consists of 9,568 observations captured every hour by some sensors located around the combined cycle power plant. All of these observations were collected over six years from 2006 to 2011 [3][4].

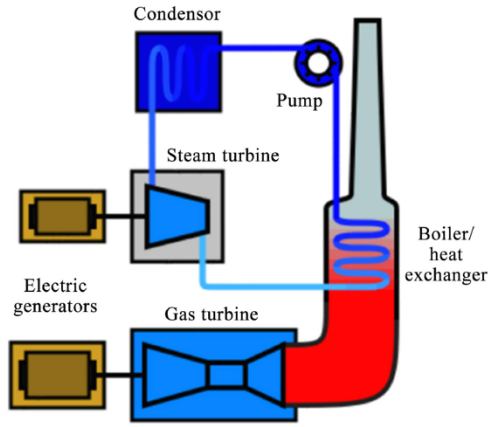

Figure 1 The combined cycle power plant diagram [2].

The aim of this paper is to examine and compare four models which are Multiple linear regression, Multilayer perceptron, k-nearest neighbor, and random forest algorithm.

The remaining of the paper is structured as follows: Section II describes the technical background. Section three discusses the dataset and features selection. Section four presents the related work. Section five explains the model design, followed by results, and discussion in Section six. Finally, the conclusion and future work section are presented in Section seven.

II. TECHNICAL BACKGROUND

A. Multiple linear Regression

Multiple linear regression is a simple linear regression with more than one input. It is an algorithm that represents the

prediction of linear relationship between multiple inputs and one output to find the best fit structure of the data based on the number of the inputs [5]–[8]. For example, finding the best plane if the inputs is two independent variable [9]. The formula of multiple linear regression is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + e_i, \quad i=1,2,\ldots,n \qquad (1)$$

y = output or dependent variable,
x_i = inputs or predictor variables,
β_0 = constant value (intercept) .
β_n = coefficient value for each input in the equation.
e_i = residuals or error of the model.

B. Multilayer perceptron

Multilayer perceptron is a type of neural networks that uses more than one perceptron to predict the output or the signals from a given input [10]–[12]. Each node represents a function it could be a step function or any other. Taking the four layers example the first layer is the inputs the last layer is the outputs and the inner layers are the hidden layers as shown in figure 2 [13]. In the perceptron algorithm, the important parameters are weight and bias. The weight changes whenever there is an error. For the training model, there are three steps. First, the model will pass the input throw the layers and that called Forward pass. Then, it will calculate the error between the desired output and the model output. Finally, the model updates the weight to match the desired one and that called backward pass [14]. Whenever there are more hidden layers the training model will be faster [13].

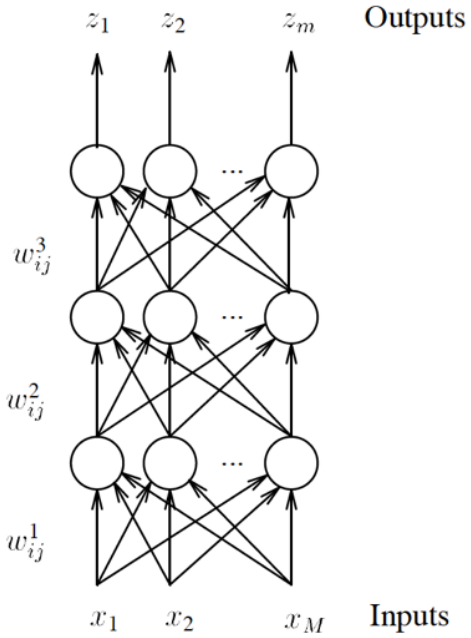

Figure 2 Multilayer Perceptron [5].

C. K-Nearest Neighbors

K-NN is a simple machine learning technique commonly used in classification problem. KNN does not build the model from the training data, that is why it is classified under lazy category [15]. KNN determines the input's class based on the number of neighbors given which is K [4]. For example, if the k = 3 and we have two classes A and B and we want to define the class of input x. First, KNN algorithm will defined the region of 3 nearest neighbors of x. the most repeated class will identify the class of x. Figure 3 shows how increasing the value of K can change the class of the input [16].

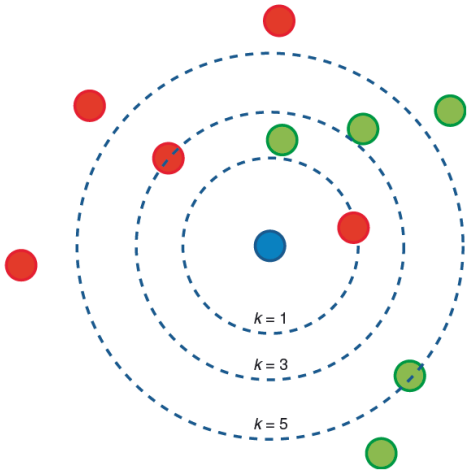

Figure 3 K-NN example [7].

D. Random Forest Algorithm

This algorithm commonly used for classification but there is an application for it in the regression filed. The random forest algorithm is a learning algorithm consists of large numbers of independent decision trees [17]–[19]. For each tree, there is a prediction for the class or the output. In each intermediate node there is a condition to specify which leaf node. The most predicted class in the decision is the output model [20]. See the figure below (a), (b), (c) and (e) trees predict the red class while the (d) tree predicts the green class. Thus, the final class is red. This prediction process goes for every single input [16].

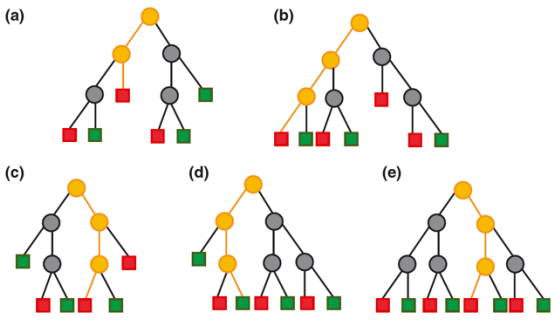

Figure 4 Random forest example [7].

E. Anderson-Darling test

Anderson Darling test is a statistical test depends on the square difference of the square of Empirical Distribution

Function tests. It is a normality test of data based on the distribution of the data [21].

III. Dataset Preprocessing and feature selections

A. Dataset Preprocessing

The Dataset discusses the electrical power that generates from the gas turbine which will affect the economy of the country. It is consisting of five attributes ambient temperature (AT), ambient pressure (AP), relative humidity (RH), exhaust vacuum (V) and electric power (EP). All these attributes related to the gas turbine situation. ambient temperature, ambient pressure, relative humidity and exhaust vacuum are inputs, and the electric power is output. There are no missing values or rows in the dataset. Each attribute shows its effects on the output so we can say all attributes or features are important.

The description for each column in the dataset:

- Ambient temperature: the temperate of the gas turbine in Celsius.
- Ambient pressure: the pressure of the gas in millibar.
- Relative humidity: ratio between the highest humidity and the current humidity in percentage.
- Exhaust vacuum: the tall of the exhaust vacuum in cm.
- Electric power: the total energy created by gas turbine in MW.

B. Outliers

There are no outliers in the dependant value of the model. Thus, there is no need to remove any value in the dataset.

C. Class of Attributes

For the regression problems we have to check the type of each feature or attribute and we have to convert the nominal variable to dummy variable for regression problem. The type is numeric for all the dependent variable and the independent variables Also, there is no need to convert any nominal value to dummy variable and there is no data matrix for the data.

D. Descriptive Summary and Normality

According to Table 1, The mean of the dependent variable is 454.365. Median equals 451.55. Mode equals 468.8. The minimum value in the dependent variable is 420.26 and the maximum is 495.76. The skewness is 0.3065 which means the dependent variable distribution is approximately symmetric because it located between 0.5 and -0.5. In addition, the

Table 1 Descriptive summary of the dependent variable.

Parameters	Values
Mean	454.365
Median	451.550
Mode	468.800
Minimum	420.260
Maximum	495.760
Skewness	0.306
Kurtosis	-1.048

kurtosis value is -1.049 which means the distribution is not normally distributed it should be 0 to be normally distributed. Q1 is 439.8, Q3 is 468.4 and the interquartile range (IQR) is 28.6 from:

$$IQR = Q3 - Q1 \quad (2)$$

For the right outliers they should be greater than 551.3. The left outliers should be less than 396.9. Applying the following equations:

$$\text{Right outlier} > Q3 + 1.5(IQR) \quad (3)$$

$$\text{Left outlier} < Q1 - 1.5(IQR) \quad (4)$$

From the above results, that is why we do not have any outliers. Moreover, the variance is 291.3 that means the data not spread out from the mean. The standard deviation is 17.1 this means there is no problem in the range of the data and the size of sample is good.

Applying Anderson-Darling normality test shows us that the dependent variable not normally distributed because the p value less than 0.05 thus, we reject the null hypothesis which is the normality. Figure 5 shows the histogram which shows that data is not normally distributed.

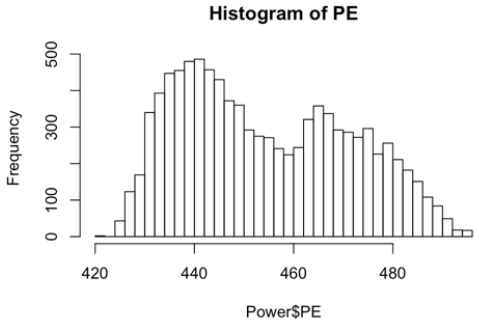

Figure 5: Histogram of the electric power variable.

E. Feature selections

Applying the stepwise selection function shows that all variables are important for the machine learning prosses because there was no discard or deletion for any variable. Moreover, the ols_ste_all_possible function shows that all variables are important for the prediction process. This function calculates the mean absolute error, R square, and more for each regression subset. The smallest MAE is 156880.41 for the subset that has all the variables. Also, it has the largest value for R square which is 0.9298 which means using this subset can predict more accurate results. Other Methods in Weka shows that deleted any variables increases mean absolute error even if the evaluator shows that one of the variables could be removed. Thus, there is no need to delete any variable.

IV. Related Work

There are several previous studies about predicting the power of combined cycle power plant using machine learning methods. One study done by Tüfekci [22] examined and

compared fifteen regression methods including simple linear regression, multiple linear regression and REPTree. The comparison based on the mean absolute error and a Root Mean-Squared Error. The best method was the Bagging method with REPTree algorithm with a mean absolute error of 2.818 and a Root Mean Squared Error of 3.787.

Moreover, Tüfekci represented a study predict the power of combined gas by comparing machine learning methods. conventional multivariate regression, additive regression, k-NN, feedforward ANN, and K-Means clustering were used to generate local and general forms. KNN best for a fine-tuned dataset. ANN yielded good results in speed and memory tests [4].

Islikaye and Cetin [23] used seven machine language methods to predict the total power of a combined cycle power plan. The best results were for K-NN, Linear Regression and RANSAC regressions.

Another study used Only ANN with random subset and different hidden layers to enhance the prediction of net power [2].

V. MODEL DESIGN

For each algorithm that I will use in this comparison, the data will be split into 80% training and 20% testing using cross-validation technique with K=10 means the training will tack 9 datasets and one dataset for the testing. Thus, there will be 9568 for the training and 1914 for the testing beads on Weka. The variance inflation factor is under 10 for all variable that means the is no multicollinearity problem exists. temperature variable has 5.92 vif, Exhaust Vacuum variable has 3.88 vif, Ambient Pressure has 1.46 vif and, Relative Humidity has 1.70 vif.

The R square of the model is 0.9298 which means that only 90% of model can be predicted. Moreover, the p-value of the model less than 0.05 as figure 13 shown. Thus, we reject the null hypothesis. That means the power is affecting by all the other attributes and it is statically significant. All the information above from the MLR model.

VI. RESULTS AND DISCUSSIONS

A. One- way Anova

One-way ANOVA uses to determine the statically significant difference between the mean of a group or more than two independent samples [24].

We have to check the normality of each sample and all of them were normally distributed because the p value is less than 0.05 so we reject the null hypothesis which is the normality. There was no statistically significant difference between the results of the algorithms because the p-value from the test is 0.0921 which is greater than 0.05 and that means do not reject the null hypothesis and there is no difference in the mean.

B. Mean Absolute Error

The multiple linear Regression shows that the MAE equals 3.6332 with reignition function equals:

Electricity Power = 454.6093 -1.9775 *Temperature - 0.2339* Exhaust Vacuum +0.0621* Ambient Pressure - 0.1581* Relative Humidity (5)

For the MLP the mean is 3.4926, KNN has 2.8725 and, the mean of random forest algorithm is 2. 3013 see figures below. From the mean results the least value of the mean is for the random forest algorithm.

Also, the random forest algorithm has the smallest value of the root mean square error with 3.3061. The MLR regression has 4.648 RMSE value. 4.4936 for the RMSE for the MLP. 4.3664 for the RMSE for the KNN.

C. Discussions

From the MAE and RMSE results we can say that the best algorithm that has the least value of the MAE and RMSE is the random forest algorithm.

VII. CONCLUSION AND FUTURE WORK

The aim of this paper was to determine the best machine learning model to predict the total electric power that could be generated by gas turbine. The random forest algorithm shows that it is the best model by having the least error.

For the future work, other algorithm could be used for the comparison such as Support Vector Machines (SVM), K star and Decision stump.

REFERENCES

[1] A. L. Polyzakis, C. Koroneos, and G. Xydis, "Optimum gas turbine cycle for combined cycle power plant," *Energy Convers. Manag.*, vol. 49, no. 4, pp. 551–563, 2008.

[2] E. A. Elfaki and A. H. Ahmed, "Prediction of Electrical Output Power of Combined Cycle Power Plant Using Regression ANN Model," *J. Power Energy Eng.*, vol. 06, no. 12, pp. 17–38, 2018.

[3] "UCI Machine Learning Repository: Combined Cycle Power Plant Data Set." .

[4] H. Kaya, P. Tüfekci, and S. F. Gürgen, "Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine," *Int. Conf. Emerg. Trends Comput. Electron. Eng. (ICETCEE 2012)*, pp. 13–18, 2012.

[5] M. Azzeh "Fuzzy Model Tree for early effort estimation," in *Proceedings - 2013 12th International Conference on Machine Learning and Applications, ICMLA 2013*, 2013, vol. 2, pp. 117–121.

[6] M. Azzeh and S. Banitaan, "Comparative analysis of soft computing techniques for predicting software effort based use case points," *IET Software*, vol. 12, no. 1. pp. 19–29, 2018.

[7] M. Azzeh, S. Banitaan, and F. Almasalha, "Pareto efficient multi-objective optimization for local tuning of analogy-based estimation," *Neural Comput. Appl.*, vol. 27, no. 8, pp. 2241–2265, 2016.

[8] M. Hosni, A. Idri, and A. Abran, "On the value of parameter tuning in heterogeneous ensembles effort estimation," *Soft Comput.*, vol. 22, no. 18, pp. 5977–6010, 2018.

[9] S. H. Brown, "Multiple Linear Regression Analysis : A Matrix Approach with MATLAB," *Alabama J. Math.*, no. Spring/Fall, pp. 1–3, 2009.

[10] I. Shahin and S. Hamsa, "Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network," *IEEE Access*, vol. 7, pp. 26777–26787, 2019.

[11] A. B. Nassif, D. Ho, and L. F. Capretz, "Towards an early software estimation using log-linear regression and a multilayer perceptron model," *J. Syst. Softw.*, vol. 86, no. 1, 2013.

[12] A. B. Nassif, L. F. Capretz, and D. Ho, "Estimating software effort using an ANN model based on use case points," in *Proceedings - 2012 11th International Conference on Machine*

Learning and Applications, ICMLA 2012, 2012, vol. 2, pp. 42–47.

[13] D. W. Ruck, S. K. Rogers, and M. Kabrisky, "Feature Selection Using a Multilayer Perceptron," *Publ. J. Neural Netw. Comput.*, vol. 2, no. 2, pp. 40–48, 1990.

[14] H. Mohamed, A. Negm, M. Zahran, and O. C. Saavedra, "Assessment of Artificial Neural Network for bathymetry estimation using High Resolution Satellite imagery in Shallow Lakes : Case Study El Burullus Lake .," *Int. Water Technol. Conf.*, no. March, pp. 434–444, 2015.

[15] M. Injadat, F. Salo, A. Essex, and A. Shami, "Bayesian Optimization with Machine Learning Algorithms Towards Anomaly Detection," *2018 IEEE Glob. Commun. Conf. GLOBECOM 2018 - Proc.*, pp. 1–6, 2018.

[16] J. B. O. Mitchell B.O., "Machine learning methods in chemoinformatics," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 4, no. 5, pp. 468–481, 2014.

[17] A. B. Nassif, "Short term power demand prediction using stochastic gradient boosting," in *International Conference on Electronic Devices, Systems, and Applications*, 2017.

[18] A. B. Nassif, L. F. Capretz, D. Ho, and M. Azzeh, "A treeboost model for software effort estimation based on use case points," in *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012*, 2012, vol. 2, pp. 314–319.

[19] A. B. Nassif, M. Azzeh, L. F. Capretz, and D. Ho, "A comparison between decision trees and decision tree forest models for software development effort estimation," in *2013 3rd International Conference on Communications and Information Technology, ICCIT 2013*, 2013, pp. 220–224.

[20] Y. L. Pavlov, "Random forests," *Random For.*, pp. 1–122, 2019.

[21] N. Mohd Razali and Y. Bee Wah, "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests," *J. Stat. Model. Anal.*, vol. 2, no. 1, pp. 21–33, 2011.

[22] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *Int. J. Electr. Power Energy Syst.*, vol. 60, pp. 126–140, 2014.

[23] A. A. Islikaye and A. Cetin, "Performance of ML methods in estimating net energy produced in a combined cycle power plant," *Proc. - 2018 6th Int. Istanbul Smart Grids Cities Congr. Fair, ICSG 2018*, pp. 217–220, 2018.

[24] E. Ostertagová and O. Ostertag, "Methodology and Application of Oneway ANOVA," *Am. J. Mech. Eng.*, vol. 1, no. 7, pp. 256–261, 2013.