In order to monitor the performance and related efficiency of a combined cycle power plant (CCPP), in addition to the best utilization of its power output, it is vital to predict its full load electrical power output. In this paper, the full load electrical power output of CCPP was predicted employing practically efficient machine learning algorithms, including linear regression, ridge regression, lasso regression, elastic net regression, random forest regression, and gradient boost regression. The original data came from an actual confidential power plant, which was working on a full load for 6 years, with four major features: ambient temperature, relative humidity, atmospheric pressure, and exhaust vacuum, and one target (electrical power output per hour). Different regression performance measures were used, including R^2 (coefficient of determination), MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and MAPE (Mean Absolute Percentage Error). Research results revealed that the gradient boost regression model outperformed other models with and without using the dimensionality reduction technique (PCA) with the highest R^2 of 0.912 and 0.872, respectively, and had the lowest MAPE of 0.872 % and 1.039 %, respectively. Moreover, prediction performance dropped slightly after using the dimensionality reduction technique almost in all regression algorithms used. The novelty in this work is summarized in predicting electrical power output in a CCPP based on a few features using simpler algorithms than reported deep learning and neural networks algorithms combined. That means a lower cost and less complicated procedure as per each, however, resulting in practically accepted results according to the evaluation metrics used

Keywords: combined cycle power plants, machine learning, predictive models, linear regression

UDC 004
DOI: 10.15587/1729-4061.2021.245663

PREDICTION OF COMBINED CYCLE POWER PLANT ELECTRICAL OUTPUT POWER USING MACHINE LEARNING REGRESSION ALGORITHMS

Nader S. Santarisi
Doctor, Associate Professor*
Sinan S. Faouri
Corresponding author
Doctor, Assistant Professor*
E-mail: s_faouri@asu.edu.jo
*Department of Mechanical and Industrial Engineering
Applied Science Private University
Al Arab str., 21, Amman, Jordan, 11931

Received date 11.10.2021
Accepted date 23.11.2021
Published date 24.12.2021

How to Cite: *Santarisi, N. S., Faouri, S. S. (2021). Prediction of combined cycle power plant electrical output power using machine learning regression algorithms. Eastern-European Journal of Enterprise Technologies, 6 (8 (114)), 16–26. doi: https://doi.org/10.15587/1729-4061.2021.245663*

1. Introduction

There are many reasons why combined cycles are more and more popular and being taken under consideration as one of the main types of power plants. The main reason is efficiency. Combined cycle power plants (CCPP) can perform more efficiently than traditional power plants by about 60 % [1]. A CCPP uses both a gas and a steam turbine together to produce up to 50 % more electricity from the same fuel than a traditional simple-cycle plant. The waste heat from the gas turbine is routed to the nearby steam turbine, which generates extra power [2]. The gas turbine compresses air and mixes it with fuel that is heated to a very high temperature. The hot air-fuel mixture moves through the gas turbine that drives an electricity generator. A heat recovery steam generator creates steam from the gas turbine exhaust heat and delivers it to the steam turbine that sends its energy to a generator where it is converted into additional electricity power output.

Predicting full load electrical power output of a base load power plant is important for accurate power production forecasts in the electrical power market to maximize the profit from the available megawatt-hours [3]. Moreover, it is a significant step toward the sustainable development of combined cycle power plants where heating load calculations are essential to optimize energy use during peak-demand seasons, and instantaneous heating loads are determined by different factors including the outdoor weather conditions [4, 5]. Furthermore, it can be integrated in a dynamic condition monitoring system in which the online performance is compared to the derived model and any deviations are diagnosed and inspected. This can ensure safe and reliable operation in various conditions [6].

Therefore, it becomes crucial to develop a method to predict power output, depending on various combinations of input/environment parameters. Several studies have been conducted in pursuit of finding accurate and efficient ways of predicting hourly electrical CCPP power output by employing different predictive models and tools (including artificial neural networks (ANN) and machine learning regressions) with corresponding accuracy performance and reliability measures (such as mean absolute error (MAE), root-mean-squared error (RMSE) and coefficient of determination R^2) as will be discussed later.

The scientific relevance is summarized in predicting electrical power output in a CCPP based on a few features using simpler algorithms than reported deep learning and neural networks algorithms combined. That means a lower cost and less complicated procedure as per each, however, resulting in practically accepted results according to the evaluation metrics used.

2. Literature review and problem statement

Machine learning is the process of equipping computers with the ability to learn by using data and experience like a human brain. The main aim of machine learning is to create models, which can train themselves to improve, perceive complex patterns, and find solutions/predictions to new problems by using the previous data [7].

Machine learning, more specifically the field of predictive modeling, is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explainability. As such, linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm [8].

The concept of linear regression was first proposed in 1894. Linear regression is a statistical test applied to a data set to define and quantify the relation between the considered variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques [9].

Dimensionality reduction is used in machine learning to avoid the curse of dimensionality and to convert the system from high to low dimension without sacrificing the important information in features. Ideally, the reduced representation should have a dimensionality that corresponds to the intrinsic dimensionality of the data [10]. The dimensionality reduction can be performed either by manually selecting the required features or by using specific techniques that reduce the system's dimension [11]. One of these techniques is Principal Component Analysis (PCA). PCA is a mathematical method that uses algorithms to reduce dimensions in a high-dimensionality system to a low dimension while keeping the maximum number of variations in the resulted features [12]. PCA works by finding directions with the highest variation of data; called principal components where working with these reduced features is much more efficient than modeling with thousands of numbers for each sample [12].

Understanding basic least squares regression is still extremely useful, but there are other improved methods that should also be considered. One issue with regular least squares is that it doesn't account for the possibility of overfitting. Ridge regression takes care of this by shrinking certain parameters. Lasso takes this step even further by allowing certain coefficients to be outright forced to zero, eliminating them from the model. Finally, Elastic Net combines the benefits of both lasso and ridge [13]. The results of [13] showed that simple least squares performed the worst on test data compared to all other models. Ridge regression provided similar results to least squares, but it did better on the test data and shrunk most of the parameters. Elastic Net ended up providing the best MSE on the test dataset by quite a wide margin. Elastic Net removed lcp, gleason and age and shrunk other parameters. Lasso also removed the consideration of age, lcp and gleason but performed slightly worse than Elastic Net.

The study in [14] aimed to develop machine learning models to accurately predict bronchiolitis severity, and to compare their predictive performance with a conventional scoring (reference) model. In a 17-center prospective study of infants (aged <1 year) hospitalized for bronchiolitis, by using routinely available pre-hospitalization data as predictors, they developed four machine learning models: Lasso regression, elastic net regression, random forest, and gradient boosted decision tree. They compared predictive models' performance with that of the reference model. The machine learning models also achieved a greater net benefit over ranges of clinical thresholds. Machine learning models consistently demonstrated a superior ability to predict acute severity and achieved greater net benefit.

Nowadays, in the context of the industrial revolution 4.0, considerable volumes of data are being generated continuously from intelligent sensors and connected objects. The proper understanding and use of these amounts of data are crucial levers of performance and innovation [15]. Machine learning is the technology that allows the full potential of big datasets to be exploited. As a branch of artificial intelligence, it enables us to discover patterns and make predictions from data based on statistics, data mining, and predictive analysis. The key goal of the study [15] was to use machine learning approaches to forecast the hourly power produced by photovoltaic panels. A comparative analysis of various predictive models including elastic net, support vector regression, random forest, and Bayesian regularized neural networks was carried out to identify the models providing the best predicting results. The principal components analysis used to reduce the dimensionality of the input data revealed six main factor components that could explain up to 91.95 % of the variation in all variables. Moreover, based on the findings of the performance metrics, it was found that non-linear models, particularly Bayesian regularized neural networks and random forest, obtained the best compromise between the predicted and observed values, with R^2=99.99 % and R^2=99.53 %, respectively, in the training phase and R^2=99.99 % and R^2=97.33 %, respectively, in the testing phase, while the lowest performance was achieved by linear models such as the elastic net algorithm with R^2=89.3 % and RMSE=0.69 kW. This is mainly because non-linear methods are better at including data dynamics and capturing non-linear correlations between variables.

In order to find accurate and efficient ways of predicting hourly electrical energy output, the researchers in [16] utilized a dataset collected over 6 years whose data points corresponded to average hourly sensor measurements when the plant was set to work with full load. The input features were ambient temperature, relative humidity and ambient pressure, which are known to be major factors in gas turbines, as well as exhaust vacuum measured from the steam turbine. They utilized conventional multivariate regression, additive regression, k-NN, feed-forward ANN and K-Means clustering to form local and global predictive models. They found that even with simple regression tools such as k-NN smoother, it is possible to predict net yield with less than 1 % relative error on average. Using more sophisticated tools and proper preprocessing it is possible to significantly increase the performance. The research in [3] explained the used methodology by: first, based on the input variables, the best subset of the dataset is explored among all feature subsets in the experiments. Then, the most successful machine learning regression method is sought for predicting full load electrical power output. Thus, the best performance of the best subset, which contains a complete set of input variables, has been observed using the most successful method with the best mean absolute error and root-mean-squared error.

Regression ANNs were used to model various systems that have high dimensionality with nonlinear relations. The system under study must have enough dataset available to train the neural network. The work in [17] was to apply and experiment with various options effects on feed-forward ANN used to obtain a regression model that predicts the electrical output power (EP) of the combined cycle power plant based on 4 inputs. The data set was obtained from an open online source. The work showed and explained the stochastic behavior of the regression neural and experiments the effect of the number of neurons of the hidden layers. A simple statistical study on the error between real values and estimated values using ANN was conducted, which showed the reliability of the model [18].

The study in [19] presented a simulation model of an existing gas-steam combined heat and power plant. The simulation models allow calculating non-measured operating parameters and energy assessment indicators. They also have the capability of adapting to the changing technical conditions of the modeled machines. Model predictive quality was verified with RMSE and R^2. The models were also used to simulate the behavior of the analyzed gas-steam plant under different operating conditions. Exemplary calculations had been presented.

In [20], electrical power output (PE) for a combined cycle gas turbine (CCGT) consisting of 9,568 data records collected over a 6-year period is evaluated by the transparent open box (TOB) machine learning method to provide accurate PE predictions and insight to prediction errors. The PE predictions derived by applying the TOB optimized data matching technique were more accurate than published predictions for the dataset from fifteen correlation-based, machine learning algorithms. Through its transparency and forensic-like auditability of its calculations for individual data records, the TOB algorithm was able to mine the dataset to provide useful insight into the interactions of the outliers with other data records. Mining the dataset also revealed significant differences in prediction accuracy achieved for different sectors of the PE distribution. This insight identified that prediction accuracy could be further improved by dividing the dataset into separately optimized subsets, three along its main PE trend plus a fourth, small subset consisting of the outliers. The TOB algorithm demonstrated its value as a machine learning tool capable of generating accurate predictions and easily auditable data mining.

The study in [21] developed a machine learning-based method to predict gas turbine performance for power generation. Two surrogate models based on high dimensional model representation and ANN were developed from real operational data to predict the operating characteristics of air compressor and turbine. Both models captured the operating characteristics well with average errors of less than 1.0 %. Since holistic ANN models have lower complexity and higher accuracy, the ANN model for predicting full-load performance was used to construct gas turbine performance correction curves.

In [22], ANNs were applied to describe the performance of a micro gas turbine. Though large, the data set did not cover the whole working range of the turbine; ANNs and an artificial neural fuzzy interference system were therefore applied to fill information gaps. The results of the investigation were also used for sensitivity analysis of the machine's behavior in different ambient conditions. ANNs can effectively evaluate both performance and emissions in real installations in any climate, the worst R^2 in the validation set was 0.9962.

In [23], a detailed investigation was aimed based on numerical thermodynamic survey and ANN modeling of the trigeneration system. The results are presented in two pivotal frameworks namely the sensitivity analysis and ANN prediction capability of proposed modeling. The underlying operative parameters were chosen as input parameters from different cycles and components, while the exergy efficiency, exergy loss, coefficient of performance, heating load exergy, lambda, gas turbine power, exergy destruction, actual outlet air compressor temperature, and heat recovery gas steam generator outlet temperature were taken as objective output parameters for the modeling purpose. It followed that multilayer perceptron neural network with back propagation algorithm resulted in the modeling reliability ranged within R^2=0.995–0.999. When the dataset treated with trainlm learning algorithm and diversified neurons, the mean squared error (MSE) was obtained equal to 0.2175.

In [24], an ANN model was constructed with the multi-layer feed-forward network type and trained with operational data using back-propagation. The results showed that the operational and performance parameters of the gas turbine can be predicted with good accuracy for varying local ambient conditions.

Different modeling techniques were proposed in [25] for determining baseline energy consumption in the industry. A combined heat and power (CHP) plant was considered in the study that was subjected to a retrofit, which consisted of the implementation of some energy-saving measures. Two different modeling methodologies were applied to the CHP plant: thermodynamic modeling and artificial neural networks (ANN). Satisfactory results are obtained with both modeling techniques. Acceptable accuracy levels of prediction were detected, confirming good capability of the models for predicting plant behavior and their suitability for baseline energy consumption determining purposes. High level of robustness was observed for ANN against uncertainty affecting measured values of variables used as input in the models.

In [6], a multi-layer perceptron (MLP) network with back propagation training was used to model the steam turbine of a combined cycle power plant with dry cooling tower. The main cooling system was modeled to predict the cooling capacity in the steam turbine exhaust using the data available to the operator. Based on that, the operators are able to predict the exhaust steam vacuum of the steam turbine (ST), which is critical in the ST output, with good accuracy. Then the data was used to predict the power output of the ST using data available to the operators through the power plant's data warehouse. It was concluded that ANN modeling is capable of predicting the electrical production of ST under varying load conditions.

The researchers in [26] argued that energy consumption has been increasing steadily due to globalization and industrialization with buildings being responsible for the biggest proportion of energy consumption. In order to control energy consumption in buildings, different policies have been proposed, from utilizing bioclimatic architectures to the use of predictive models within control approaches. There are mainly three groups of predictive models including engineering, statistical and artificial intelligence models. The main objective of the study was to compare a neural network model, which was designed utilizing statistical and analytical methods, with a group of neural network

models designed benefiting from a multi-objective genetic algorithm (MOGA). Moreover, the neural network models were compared to a naive autoregressive baseline model. The models were intended to predict electric power demand at the Solar Energy Research Center bioclimatic building located at the University of Almeria, Spain. Experimental results showed that the models obtained from MOGA perform comparably to the model obtained through a statistical and analytical approach, but they use only 0.8 % of data samples and have lower model complexity.

Heat rate of a combined cycle power plant is a parameter that is typically used to assess how efficient a power plant is. In [27], the CCPP heat rate was predicted using an ANN method to support the maintenance team in monitoring the efficiency of the CCPP. The ANN method used fuel gas heat input, CO_2 percentage, and power output as input parameters. Approximately 4322 actual operation data are generated from the digital control system (DCS) in a year. These data were used for ANN training and prediction. Seven parameter variations were developed to find the best parameter variation to predict heat rate. The ANN model that utilized three parameters as input data had the best prediction heat rate data with a regression coefficient of determination value of 0.995.

The above literature demonstrated the level of details and related modeling and computational complexity to predict performance (related output) that highlighted the need for testing the performance of kind-of-simpler algorithms, using a few features, than reported deep learning and neural networks algorithms. These would be alternative algorithms to overcome these difficulties at lower cost and less complicated procedures, while resulting in practically accepted results according to the known used evaluation performance metrics.

3. The aim and objectives of the study

The aim of the study is to predict the full load electrical power output of CCPP depending on four features: ambient temperature, relative humidity, atmospheric pressure, and exhaust vacuum, and one target (electrical power output per hour).

To achieve this aim, the following objectives are accomplished:

– to make sure the data is clean with no missing values, duplications or outliers;

– to implement the following models: linear regression, ridge regression, lasso regression, elastic net regression, random forest regression, and gradient boost regression then evaluate the models using R^2, MAE, MSE, RMSE, MAPE;

– to use the dimensionality reduction technique (PCA) and re-do the point (2) in objectives, then compare results with and without using dimensionality reduction and see where the best results occur.

4. Materials and methods

4. 1. Data origin and information

The original data came from a real confidential power plant, which was working on a full load for 6 years. The data has been downloaded from [28]. It consists of 9,568 instances with four features (ambient temperature (AT), relative humidity (RH), atmospheric pressure (AP), exhaust vacuum (V)) and one target (electrical power output per hour (PE)). The downloaded data was clean with no missing values, thus, no need for any imputation technique before the modeling phase. It has been noted also that there are no obscure outliers in the data, which made the correlation procedure between features and target quite straightforward.

4. 2. Choice of learning algorithms

Choosing the right algorithm to train data on is not an easy task. However, the obvious correlation between PE and AT seen in Fig. 2 excluded our assumptions to specific linear originated models. The use of ensemble algorithms should result in more accurate performance theoretically, because it combines a set of weak learners together. The chosen algorithms are linear regression, ridge regression, lasso regression, elastic net regression, random forest regression and gradient boost regression.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications [4]. This is because models that depend linearly on their unknown parameters are easier to fit than models that are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable [29].

Linear regression models are often fitted using the least-squares approach; by fitting the least square residuals between the observed dataset points and predicted dataset points. But they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least-squares cost function as in ridge regression (L2-norm penalty) and lasso regression (L1-norm penalty) [29].

Ridge Regression, also called Tikhonov regularization is a regularized version of the Linear Regression: adding a regularization term, commonly referred to as alpha, the cost function, the learning algorithm is forced to keep the weight as low as possible. It adds a detractor (penalty factor) to the cost function. This determines the loss of importance of the value (check) of a feature, which, depending on the penalty, can be more or less accentuated. The strength of the penalty is tunable controlled, that is, by a hyperparameter that must be set. Speaking of regularization in general, there are two types of penalties [30]:

– $L1$ (absolute size) penalizes the absolute value of the model coefficient 0073;

– $L2$ (squared size) penalizes the square of the value of the model coefficients.

Ridge Regression uses the $L2$ penalty. In practice, this produces small coefficients, but none of them are ever canceled out. Therefore, the coefficients are never 0. The phenomenon is called feature shrinkage [30].

The word "LASSO" stands for Least Absolute Shrinkage and Selection Operator. It is a statistical formula for the regularization of data models and feature selection. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i. e., models with fewer parameters). This particular type of regression is

well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. Lasso Regression uses the $L1$ regularization technique. It is used when we have a greater number of features because it automatically performs feature selection [31].

Elastic net is a popular type of regularized linear regression that combines two popular penalties, specifically the $L1$ and $L2$ penalty functions. A hyperparameter, alpha "α", is provided to assign how much weight is given to each of the $L1$ and $L2$ penalties. Alpha is a value between 0 and 1 and is used to weight the contribution of the $L1$ penalty and "one minus the alpha" value is used to weight the $L2$ penalty as shown in (1) [32].

$$\text{Elastic net penalty}=(\alpha(\text{L1\_penalty}))+ +((1-\alpha)(\text{L2\_penalty})). \quad (1)$$

The benefit is that elastic net allows a balance of both penalties, which can result in better performance than a model with either one or the other penalty on some problems.

Another hyperparameter is provided called lambda "λ" that controls the weighting of the sum of both penalties to the loss function as shown in (2). A default value of 1.0 is used to use the fully weighted penalty; a value of 0 excludes the penalty. Very small values of lambada, such as 0.001 or smaller, are common [32]. Elastic net performs better in a large dataset.

$$\text{Elastic\_net\_loss}=\text{loss}+(\lambda(\text{elastic\_net\_penalty})). \quad (2)$$

An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. Random forest is a supervised learning algorithm, which uses the ensemble learning method for classification and regression. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forest is one of the most accurate learning algorithms available. It can handle thousands of input variables without variable deletion and gives estimates of what variables are important in the classification. Furthermore, it generates an internal unbiased estimate of the generalization error as the forest building progresses. Meanwhile, for data including categorical variables with different numbers of levels, random forests are biased in favor of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable for this type of data [33].

Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modeling problems. It is also known as gradient tree boosting, stochastic gradient boosting (an extension), and gradient boosting machines, or GBM for short. Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This is a type of ensemble machine learning model referred to as boosting [34].

Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, "gradient boosting," as the loss gradient is minimized as the model is fit, much like a neural network. Gradient boosting performs well, if not the best, on a wide range of tabular datasets, and versions of the algorithm like XGBoost and LightBoost often play an important role in winning machine learning competitions [34].

4. 3. Evaluation metrics

Deciding which evaluation metrics to use in order to evaluate training performance isn't an easy task due to possible data imbalance, for example. Regarding the problem in this work, as can be noted clearly in Fig. 2, the linearity of the regression leads us to choose the following evaluation measures: R^2 (coefficient of determination), MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and MAPE (Mean Absolute Percentage Error). MAE calculates the absolute difference between actual and predicted values (3). MSE calculates the squared error or distance between actual and predicted values (4). The reason behind squaring values is to cancel the negativity effect. One con of this method is that it penalizes the outliers when it squares the outputs, which isn't the case in MAE. Another evaluation metric used in this work is RMSE (5). As its name suggests, it roots down the value of MSE, which makes the output more interpretable and has the same intended unit. R^2 is also used in this study (6), which is quite different from the latter evaluation measures in that it doesn't measure the loss in a sense, instead, it measures how well the model performs. It's called coefficient of determination and the higher its value is the better. It takes values between 0 and 1. The last evaluation metric in this work is MAPE, which measures the percentage accuracy of predicted outputs to the actual outputs with a percentage error (7).

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}\right|, \quad (3)$$

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}\right)^2, \quad (4)$$

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}\right)^2}, \quad (5)$$

$$R^2=\frac{\text{MSS}}{\text{TSS}}=\frac{\text{TSS}-\text{RSS}}{\text{TSS}}, \quad (6)$$

$$\text{MAPE}=\frac{100}{n}\sum_{i=1}^{n}\left|\frac{\left(y_i-\hat{y}\right)}{y_i}\right|, \quad (7)$$

where y_i – the actual i^{th} output, $\hat{y}$ – the predicted output, n – number of instances, MSS is the model sum of squares (also known as ESS, or explained sum of squares), which is the sum of the squares of the prediction from the linear regression minus the mean for that variable; TSS is the total sum of squares associated with the outcome variable, which is the sum of the squares of the measurements minus their mean; and RSS is the residual sum of squares, which is the sum of the squares of the measurements minus the prediction from the linear regression [35, 36]. That means:

$$\text{TSS: Total Sum of Squares}=\sum_{i=1}^{n}\left(y_i-y_m\right)^2,$$

$$\text{MSS: Model Sum of Squares}=\sum_{i=1}^{n}\left(\hat{y}-y_m\right)^2,$$

$$\text{RSS: Residual Sum of Squares}=\sum_{i=1}^{n}\left(y_i-\hat{y}\right),$$

where y_m is the mean value of the output.

5. Results of research of predicting the full load electrical power output of CCPP depending on four features

5. 1. Exploratory data analysis (EDA)

Exploratory data analysis resulted in clean data by the means of no missing values, outliers or duplications in our dataset according to the output of our code. This pathed the way for the next objective of modeling. The scatter matrix in Fig. 1 shows an obvious linear relationship between AT and PE.

The PE vs AT relationship is enlarged for a better visuality in Fig. 2. As shown in Fig. 2, a direct and linear relationship exists between PE and AT.

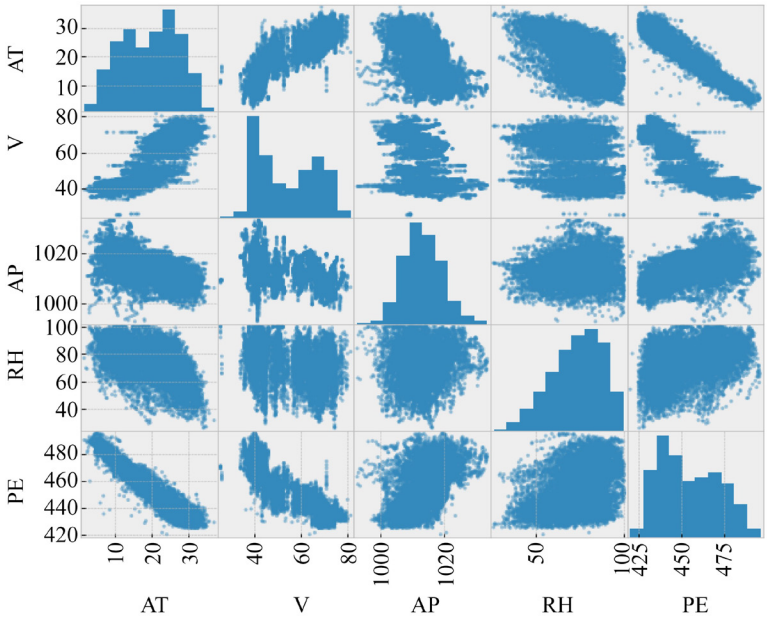

Fig. 1. Scatter matrix showing correlation between features and target: AT – ambient temperature; RH – relative humidity; AP – atmospheric pressure; V – exhaust vacuum; PE – electrical power output

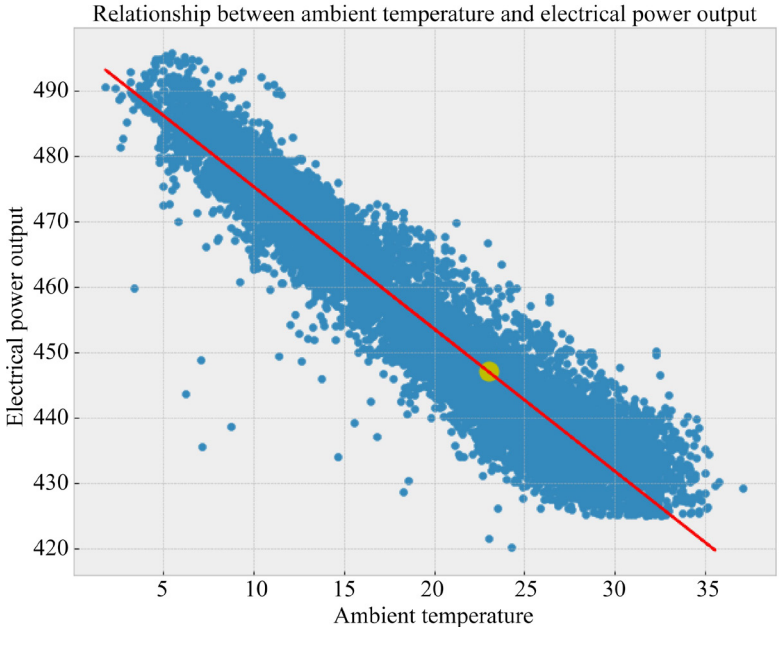

Fig. 2. Correlation between ambient temperature and electrical power output per hour

When AT increases, PE decreases. The red line is drawn from the linear regression model (discussed) and the yellow dot represents a prediction that follows this linearity where PE is 447.08 kWh when AT is 23 °C.

5. 2. Results of implementing models without dimensionality reduction

This section includes results obtained from training CCPP data and measuring its performance without using feature reduction. As shown in Fig. 2, there's clearly a linear relationship between AT and PE as mentioned in Section 5. 1. According to this linearity, in our first set of results we had AT as our feature data set and PE as our label then trained the data on six different models, where four of them are linear originated estimators (linear regression, ridge regression, lasso regression, elastic net) and the other two are ensemble methods (random forest and gradient boost models) that depend on a combination of weak algorithms then evaluated their performance using five different evaluation metrics as shown in Table 1. It's worth noting that results in Table 1 are performed without using a feature reduction method. Moreover, we do not have to manually set the values of the regularization parameters (α and λ) in our modeling procedure, as they are optimized automatically in the models used.

As shown in Table 1, for the ensemble methods (random forest and gradient boost regressors), random forest regression performed worse than linear originated methods while gradient boost regressor was the best model performer between all 6 models with R^2 of 0.912 and the lowest MAPE of 0.872 %. The high performance of gradient boost regressor is mainly because non-linear methods are better at including data dynamics and capturing non-linear correlations between variables, while the low performance of random forest regression can be attributed to the expected variables' different number of levels as random forests are biased in favor of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable in such conditions. Meanwhile, all four linear originated models performed in a very similar way, which can be explained as a result of the used modeling variables and the related linearity assumption. The details of achieved results are discussed.

The R^2 value for the six different regression models is shown clearly in Fig. 3.

Fig. 4 shows MAE for the six different models. Again, it's noted here that the linear originated models perform similarly with an MAE of around 4.27 approximately.

Fig. 5 shows the MSE of the six regression models.

Fig. 6 shows the RMSE value of the six models.

MAPE of the six models is shown in Fig. 7.

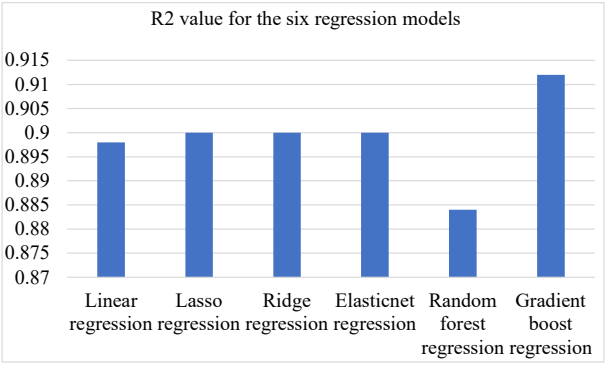

Fig. 3. R^2 value for the six regression models without using PCA

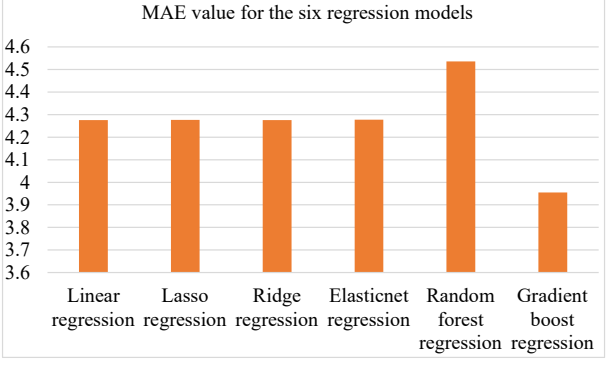

Fig. 4. MAE value for the six regression models without using PCA

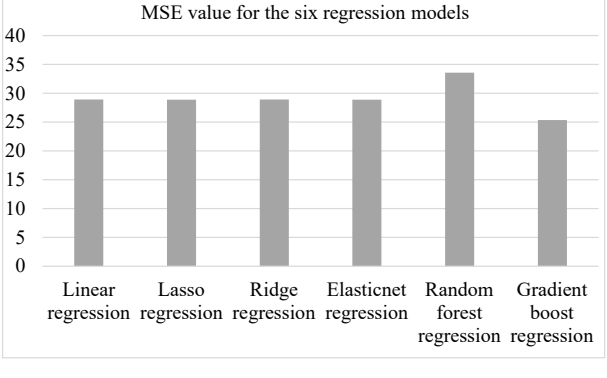

Fig. 5. MSE value for the six regression models without using PCA

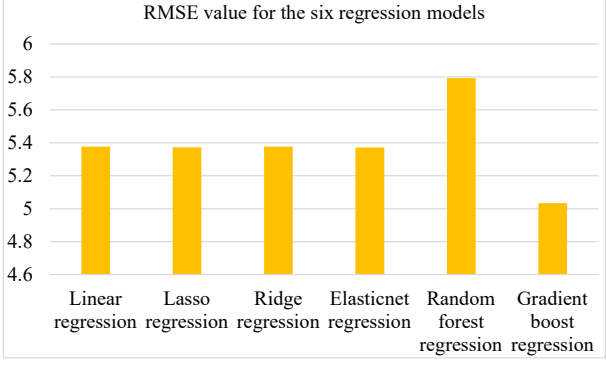

Fig. 6. RMSE value for the six regression models without using PCA

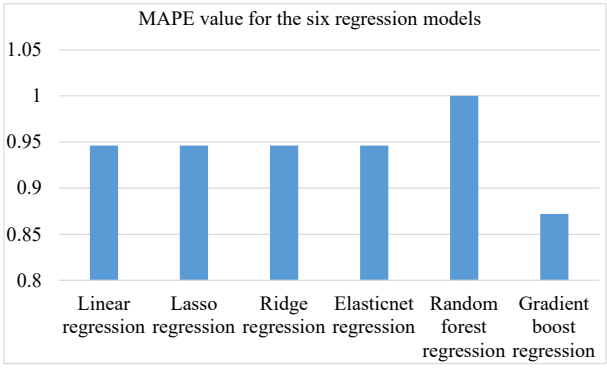

Fig. 7. MAPE value for the six regression models without using PCA

Table 1

Comparison of models' evaluation metrics without using dimensionality reduction

Evaluation metric	Linear regression	Lasso regression	Ridge regression	Elastic net regression	Random forest regression	Gradient boost regression
R^2	0.898	0.900	0.900	0.900	0.884	0.912
MAE	4.276	4.277	4.276	4.278	4.536	3.955
MSE	28.912	28.872	28.912	28.866	33.569	25.344
RMSE	5.376	5.373	5.376	5.372	5.793	5.034
MAPE (%)	0.946	0.946	0.946	0.946	1.000	0.872

5. 3. Results of implementing models using dimensionality reduction

Table 2 shows the performance metrics after using the dimensionality reduction technique, Principal Component Analysis (PCA), which aims to select features with the lowest loss.

Table 2

Comparison of models' evaluation metrics with using dimensionality reduction technique (PCA)

Evaluation metric	Linear regression	Lasso regression	Ridge regression	Elastic net regression	Random forest regression	Gradient boost regression
R^2	0.851	0.850	0.851	0.851	0.862	0.872
MAE	5.134	5.135	5.134	5.135	4.902	4.742
MSE	43.201	43.219	43.201	43.215	39.869	37.020
RMSE	6.572	6.574	6.572	6.573	6.314	6.084
MAPE (%)	1.126	1.126	1.126	1.126	1.073	1.039

As shown in Table 2, again, gradient boost regressor performs better than the other ensemble method; random forest regressor, and even better than the other four linear originated models with the highest R^2 of 0.872 and lowest MAPE of 1.039 %. However, this time after using PCA, it is noticed that random forest regressor isn't the worst model in this group. Actually, random forest regressor performance results outperform all the four linear originated models in terms of R^2 (0.862), MAE (4.902), MSE (39.869), RMSE (6.314) and MAPE of 1.073 %. The high performance of gradient boost regressor and random forest is mainly because non-linear methods are better at including data dynamics and capturing non-linear correlations between variables. However, the performance metrics of random forest regres-

sor after using PCA are still behind the performance results of gradient boost regressor. The details of achieved results are discussed.

Fig. 8 shows the performance of the six models after applying the dimensionality reduction technique of PCA measured by the evaluation metric R^2.

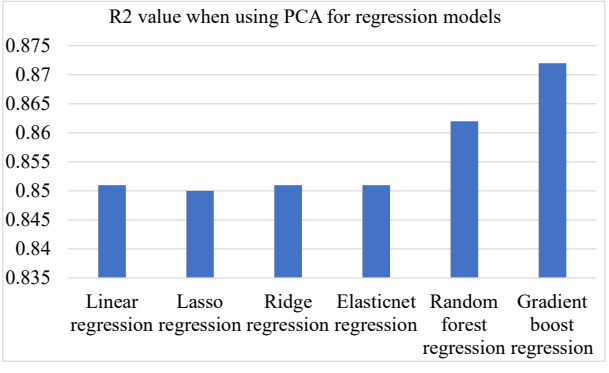

Fig. 8. R^2 value for the six regression models when using PCA

Fig. 9 shows MAE of the six models in this study after applying PCA.

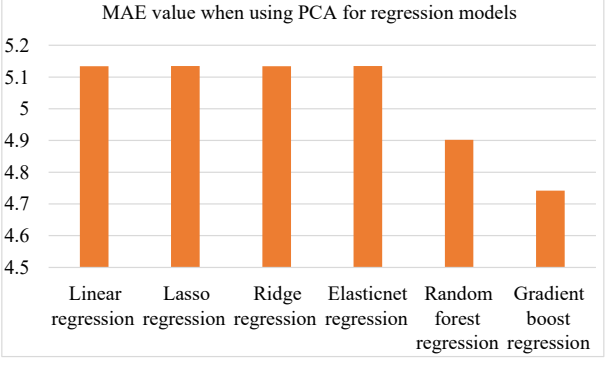

Fig. 9. MAE value for the six regression models when using PCA

Fig. 10 shows a very close performance of the four linear originated models when applying PCA and measuring it using MSE.

Fig. 11 shows the performance of the six models for the evaluation metric of RMSE when using the dimensionality reduction technique of PCA.

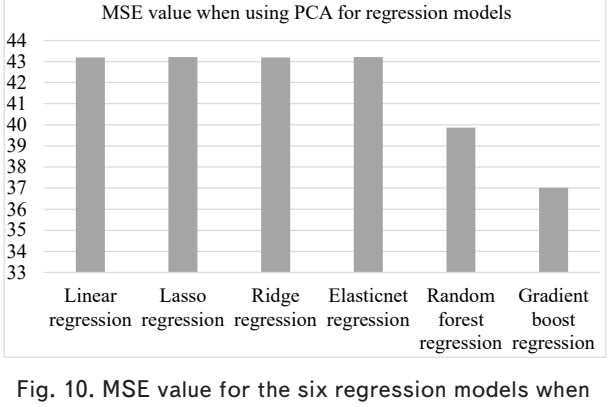

Fig. 10. MSE value for the six regression models when using PCA

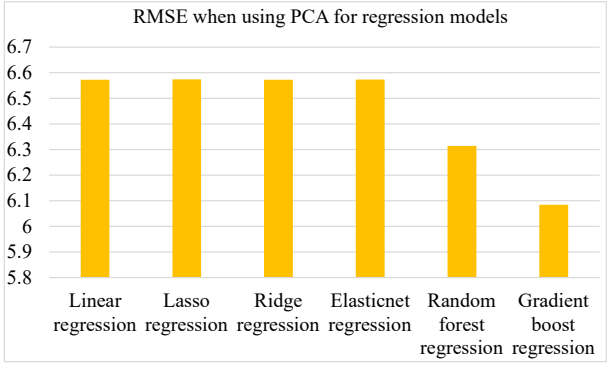

Fig. 11. RMSE value for the six regression models when using PCA

Fig. 12 shows the performance of the six models after applying the evaluation metric of MAPE when using feature selection of PCA.

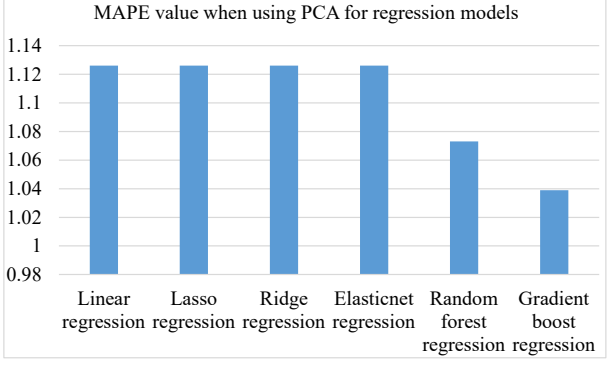

Fig. 12. MAPE value for the six regression models when using PCA

6. Discussion of results of predicting the full load electrical power

According to proven linearity between AT and PE, the first set of results, shown in Table 1, where six different models: four of them are linear-originated estimators (linear regression, ridge regression, lasso regression, elastic net) and the other two are ensemble methods (random forest and gradient boost models), without using a feature reduction method. The following details the discussion of the performance metrics results found in Table 1.

It can be noted in Fig. 3 that almost all linear originated estimators (linear regression, lasso regression, ridge regression, elastic net regression) perform similarly for this evaluation metric. As for the ensemble estimators, it can be shown that random forest regressor performs worse than the latter four linear originated models with an R^2 value of 0.884 while the other ensemble model (gradient boost regressor) performs better than the other five models in Fig. 3 with an R^2 of 0.912.

The random forest regressor model performs worse than the linear originated model with an MAE of 4.536. However, gradient boost regressor as an ensemble method performs better than the other five models with an MAE of 3.955.

The first four linear originated models perform almost similarly with a very close value of MSE. It can be noted clearly from Fig. 5 that random forest regressor performs worse than linear originated models with an MSE of 33.569.

Our second ensemble method, gradient boost regressor, again performs the best between the previous five models in this evaluation metric with an MSE of 25.344.

The linear originated models perform similarly with an RMSE of around 5.37. Random forest regressor performs worse in this evaluation metric with an RMSE of 5.793. The best model performance of RMSE is again gradient boost regressor with a value of 5.034 for this evaluation metric.

The linear originated models produced a good performance in this evaluation metric with an MAPE of 0.946 %, which indicates a high level of accuracy. Random forest regressor performs a little bit worse with an MAPE of 1.00 %. The best performance for this evaluation metric goes to the second ensemble method (gradient boost regressor) with an MAPE of 0.872 %.

Moreover, Table 2 shows the performance metrics after using the dimensionality reduction technique, Principal Component Analysis (PCA), with the same assumptions considered for Table 1 results. The following details the discussion of the performance metrics results found in Table 2.

The four linear originated models perform similarly for this evaluation metric with an R^2 of around 0.85. Random forest regressor clearly performs a little bit better than the other four linear originated models after applying PCA. Gradient boost regressor has the best performance for R^2 after applying the dimensionality reduction technique with a value of 0.872.

It can be clearly noted that the four linear originated models perform similarly with an MAE of around 5.13. The ensemble methods both perform better than the four linear originated models with a value of 4.902 for random forest regressor and best performance of 4.742 for gradient boost regressor for this evaluation metric after using PCA.

Random forest regressor performs better than the latter four linear originated models for this evaluation metric with an MSE of 39.869. Again, gradient boost regressor performs the best in MSE measurement with a value of 37.02 after applying PCA.

The four linear originated models perform similarly with an RMSE of around 6.57. The ensemble methods perform better than the previous four linear originated models with RMSE values of 6.314 and 6.084 for random forest regressor and gradient boost regressor as a best performer, respectively.

The four linear originated models perform similarly with an MAPE value of 1.126 % when using PCA. The performance improves when using random forest regressor with an MAPE of 1.073 %. The best performance of MAPE after applying the dimensionality reduction technique of PCA goes to gradient boost regressor with an MAPE of 1.039 %.

Comparing the achieved results shown in Tables 1, 2 reveals that performance dropped after using the dimensionality reduction technique (PCA) almost in all evaluation metrics that were used. The dimensionality reduction technique usually reduces complexity and cost in the system while retaining the most important information in the features. Modeling two features instead of one linearly related feature with the target participated in the slight drop in performance. Moreover, despite the kind-of-simplicity of the techniques used (with and without using the dimensionality reduction technique) compared to more detailed and complicated techniques found in literature, they provided practically accepted results that confirms according to performance metrics and close to other's outcomes.

One of the limitations of this study is that if we have a much smaller dataset, it is not guaranteed that the performance of the chosen models will stay the same as it is known that machine learning models require a large amount of dataset to increase the performance. Also, the data that we have worked with is considered clean with no missing values, duplicates or outliers. Having any of the former circumstances may change the performance of the chosen models.

7. Conclusions

1. Using original data, came from a real confidential power plant, which was working on a full load for 6 years, we used machine learning models to predict CCPP full load electrical power output per hour depending on four main features (AT, RH, V, and AP). It has turned out that the data is clean with no missing values, duplicates or outliers.

2. It has been revealed that the gradient boost regression model outperformed linear regression, ridge regression, lasso regression, elastic net regression, and random forest regression, when using the dimensionality reduction technique (PCA) with the highest R^2 of 0.912 and the lowest MAPE of 0.872 %. It has been revealed that the gradient boost regression model outperformed linear regression, ridge regression, lasso regression, elastic net regression, and random forest regression, without using the dimensionality reduction technique (PCA) with the highest R^2 of 0.872, and the lowest MAPE of 1.039 %.

3. Moreover, prediction performance dropped slightly after using the dimensionality reduction technique almost in all regression algorithms used. The research results aligned with similar research, while MAPE was an added-used performance measure in this field. Finally, we were able to predict electrical power output in a CCPP based on a few features using simpler algorithms than reported deep learning and neural networks algorithms combined. That means a lower cost and less complicated procedure as per each, however, resulting in practically accepted results according to the evaluation metrics used. As a future work, this study can be extended by implementing other algorithms and testing on different kinds of power plants.

Acknowledgments

The authors would like to thank the students Rami Sayoori, Mousa Tawasha, Ayham Bushnaq, and Mohammad Alshanawani for their related assistance to this study.

References

1. Hoang, T.-D., Pawluskiewicz, D. K. (2016). The efficiency analysis of different combined cycle power plants based on the impact of selected parameters. International Journal of Smart Grid and Clean Energy, 5 (2), 77–85. doi: https://doi.org/10.12720/sgce.5.2.77-85
2. Combined cycle power plant: how it works. Available at: https://www.ge.com/gas-power/resources/education/combined-cycle-power-plants

3. Tüfekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. International Journal of Electrical Power & Energy Systems, 60, 126–140. doi: https://doi.org/10.1016/j.ijepes.2014.02.027
4. Moayedi, H., Mosavi, A. (2021). Electrical Power Prediction through a Combination of Multilayer Perceptron with Water Cycle Ant Lion and Satin Bowerbird Searching Optimizers. Sustainability, 13 (4), 2336. doi: https://doi.org/10.3390/su13042336
5. Sholahudin, S., Han, H. (2015). Heating Load Predictions using The Static Neural Networks Method. International Journal of Technology, 6 (6), 946. doi: https://doi.org/10.14716/ijtech.v6i6.1902
6. Dehghani Samani, A. (2018). Combined cycle power plant with indirect dry cooling tower forecasting using artificial neural network. Decision Science Letters, 7, 131–142. doi: https://doi.org/10.5267/j.dsl.2017.6.004
7. Çelik, Ö. (2018). A Research on Machine Learning Methods and Its Applications. Journal of Educational Technology and Online Learning, 1 (3), 25–40. doi: https://doi.org/10.31681/jetol.457046
8. Brownlee, J. (2016). Linear Regression for Machine Learning. Machine Learning Algorithms. Available at: https://machinelearningmastery.com/linear-regression-for-machine-learning/
9. Kumari, K., Yadav, S. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences, 4 (1), 33. doi: https://doi.org/10.4103/jpcs.jpcs_8_18
10. Van Der Maaten, L., Postma, E., van den Herik, J. (2009). Dimensionality Reduction: A Comparative Review. Available at: https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf
11. Mladenić, D. (2006). Feature Selection for Dimensionality Reduction. Lecture Notes in Computer Science, 84–102. doi: https://doi.org/10.1007/11752790_5
12. Ringnér, M. (2008). What is principal component analysis? Nature Biotechnology, 26 (3), 303–304. doi: https://doi.org/10.1038/nbt0308-303
13. Sneiderman, R. (2020). From Linear Regression to Ridge Regression, the Lasso, and the Elastic Net. And why you should learn alternative regression techniques. Available at: https://towardsdatascience.com/from-linear-regression-to-ridge-regression-the-lasso-and-the-elastic-net-4eaecaf5f7e6
14. Raita, Y., Camargo, C. A., Macias, C. G., Mansbach, J. M., Piedra, P. A., Porter, S. C. et. al. (2020). Machine learning-based prediction of acute severity in infants hospitalized for bronchiolitis: a multicenter prospective study. Scientific Reports, 10 (1). doi: https://doi.org/10.1038/s41598-020-67629-8
15. Chahboun, S., Maaroufi, M. (2021). Principal Component Analysis and Machine Learning Approaches for Photovoltaic Power Prediction: A Comparative Study. Applied Sciences, 11 (17), 7943. doi: https://doi.org/10.3390/app11177943
16. Kaya, H., Tüfekci, P., Gürgen, S. F. (2012). Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine. International Conference on Emerging Trends in Computer and Electronics Engineering (ICETCEE'2012), 13–18. Available at: http://psrcentre.org/images/extraimages/70.%20312595.pdf
17. Elfaki, E., Hassan, A. H. A. (2018). Prediction of Electrical Output Power of Combined Cycle Power Plant Using Regression ANN Model. International Journal of Computer Science and Control Engineering, 6 (2), 9–21. Available at: https://zenodo.org/record/1285164#.YaX5l1VByUk
18. Elfaki, E. A., Ahmed, A. H. (2018). Prediction of Electrical Output Power of Combined Cycle Power Plant Using Regression ANN Model. Journal of Power and Energy Engineering, 06 (12), 17–38. doi: https://doi.org/10.4236/jpee.2018.612002
19. Plis, M., Rusinowski, H. (2018). A mathematical model of an existing gas-steam combined heat and power plant for thermal diagnostic systems. Energy, 156, 606–619. doi: https://doi.org/10.1016/j.energy.2018.05.113
20. Wood, D. A. (2020). Combined cycle gas turbine power output prediction and data mining with optimized data matching algorithm. SN Applied Sciences, 2 (3). doi: https://doi.org/10.1007/s42452-020-2249-7
21. Liu, Z., Karimi, I. A. (2020). Gas turbine performance prediction via machine learning. Energy, 192, 116627. doi: https://doi.org/10.1016/j.energy.2019.116627
22. Bartolini, C. M., Caresana, F., Comodi, G., Pelagalli, L., Renzi, M., Vagni, S. (2011). Application of artificial neural networks to micro gas turbines. Energy Conversion and Management, 52 (1), 781–788. doi: https://doi.org/10.1016/j.enconman.2010.08.003
23. Anvari, S., Taghavifar, H., Saray, R. K., Khalilarya, S., Jafarmadar, S. (2015). Implementation of ANN on CCHP system to predict trigeneration performance with consideration of various operative factors. Energy Conversion and Management, 101, 503–514. doi: https://doi.org/10.1016/j.enconman.2015.05.045
24. Fast, M., Assadi, M., De, S. (2009). Development and multi-utility of an ANN model for an industrial gas turbine. Applied Energy, 86 (1), 9–17. doi: https://doi.org/10.1016/j.apenergy.2008.03.018
25. Rossi, F., Velázquez, D., Monedero, I., Biscarri, F. (2014). Artificial neural networks and physical modeling for determination of baseline consumption of CHP plants. Expert Systems with Applications, 41 (10), 4658–4669. doi: https://doi.org/10.1016/j.eswa.2014.02.001

26. Khosravani, H., Castilla, M., Berenguel, M., Ruano, A., Ferreira, P. (2016). A Comparison of Energy Consumption Prediction Models Based on Neural Networks of a Bioclimatic Building. Energies, 9 (1), 57. doi: https://doi.org/10.3390/en9010057
27. Arferiandi, Y. D., Caesarendra, W., Nugraha, H. (2021). Heat Rate Prediction of Combined Cycle Power Plant Using an Artificial Neural Network (ANN) Method. Sensors, 21 (4), 1022. doi: https://doi.org/10.3390/s21041022
28. Kaggle. Available at: https://www.kaggle.com/gova26/airpressure
29. Linear regression. Wikipedia. Available at: https://en.wikipedia.org/wiki/Linear_regression
30. Ridge Regression. Available at: https://andreaprovino.it/ridge-regression/
31. A Complete understanding of LASSO Regression (2020). Available at: https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/
32. Brownlee, J. (2020). How to Develop Elastic Net Regression Models in Python. Python Machine Learning. Available at: https://machinelearningmastery.com/elastic-net-regression-in-python/
33. Chakure, A. (2019). Random Forest Regression. Available at: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f
34. Brownlee, J. (2020). How to Develop a Gradient Boosting Machine Ensemble in Python. Ensemble Learning. Available at: https://machinelearningmastery.com/gradient-boosting-machine-ensemble-in-python/
35. Thakur, M. Coefficient of Determination Formula. Available at: https://www.educba.com/coefficient-of-determination-formula/
36. Enders, F. B. Coefficient of determination. Available at: https://www.britannica.com/science/coefficient-of-determination