International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Optimizing Hourly Power Output Forecasting in Combined Cycle Power Plants Using Machine Learning Regression Techniques

Muhammad Ali Asad [1], *Tallat Mahmood* [2], *Inayat Ullah Jamal Khan* [3], *Shajjad Hossen* [4], *Haque Medha* [5]

[1*] School of Computer Sciences and Engineering, National Textile University, FSD Pakistan, Email: asadali2000@126.com

[2*] School of Computer Science and Artificial Intelligence, Wuhan Textile University, China Email: chtallat1@gmail.com

[3,4,5*] School of Electrical Engineering and Automation, Jiangsu Normal University, China Email: inayatst1234321@gmail.com

ABSTRACT

Optimizing the available megawatt hours in baseload power plants requires precise fore- casting of full-load electrical output. This study evaluates various machine learning regression techniques to develop a robust model for predicting hourly output in a combined cycle power plant. The dataset comprises 9,568 data points spanning the years 2006 to 2011, collected from a high-efficiency operation.

Key variables influencing the power plant's performance include ambient temperature,

atmospheric pressure, relative humidity, and exhaust steam pressure. The target variable, elec- trical power output, varies significantly based on these environmental factors. After rigorous evaluation of multiple regression approaches, the ensemble method employing a bagged tree approach with the full set of input variables emerged as the top performer.

This model achieved a commendable root mean square error (RMSE) of 3.51 and mean absolute error (MAE) of 2.5484 during validation, indicating its superior predictive accuracy. Notably, the ensemble method also demonstrated impressive computational efficiency, capable of processing approximately 30,000 observations per second. These findings underscore the effectiveness of advanced machine learning techniques in optimizing operational forecasting for enhanced efficiency and reliability in power generation.

Keywords: Machine learning, Regression, Optimal Power Forecasting

1. Introduction

Solving several nonlinear equations is complicated, thermodynamic approaches to study- ing systems sometimes depend on assumptions. These presumptions are important because they consider how unpredictable solutions can be. Without them, it would be very challenging and time-consuming to solve these equations in order to investigate real-world applications. Ma- chine learning approaches are widely employed in place of thermodynamic ones to get over this problem, particularly in systems with random inputs and outputs. For example[1], uses an Artificial Neural Network (ANN) model based on real plant data to investigate how variables like atmospheric pressure, temperature, humidity, wind speed, and direction affect power plant performance.

Several machine learning regression algorithms are applied to investigate a particular ther- modynamic system the combined cycle power plant (CCPP). Heat recovery systems, steam tur- bines, and gas turbines are some of the parts of this facility [3]. For a plant of this type to operate profitably and efficiently, it is essential to predict its electrical production. 9,568 data points were gathered over the course of six years (2006–2011) from a CCPP that was working at maximum capacity [4]. This dataset was used in the study. The net hourly electrical energy output (EP) of the plant is predicted using variables like exhaust vacuum (V), relative humidity (RH), ambient pressure (AP), and temperature (T). The purpose of this study is to assess the predictive accuracy of several regression techniques for the full-load electrical power output of a base load-operated CCPP.

This study aims to evaluate the prediction accuracy of several regression techniques in predicting the electrical power output of a base load-operated CCPP at full load. This paper's second section describes the approaches used and how machine learning techniques were applied in this thermodynamic setting. The actual findings from these analyses are presented in Section 3, which provides insight into how well various regression models predict CCPP output. The results are contextualized and critically analyzed in Section 4, along with the study's implications and key takeaways. Section 5 concludes the study by summarizing the main discoveries and outlining potential directions for more research in this area.

2. Methods and materials

The 480 MW designed capacity of the combined cycle power plant (CCPP) utilized in this study is made up of two 160 MW ABB 13E2 gas turbines, two dual-pressure HRSGs (heat re- covery steam generators), and one 160 MW ABB steam turbine. The operation of the gas and steam turbines is greatly influenced by several environmental factors, including ambient tem- perature (AT), atmospheric pressure (AP), relative humidity (RH), and exhaust steam pressure

(V). The target variable in this dataset is the electrical power output produced by both types of turbines, with the ambient and steam parameters being considered as input variables. Certain ranges for every variable are reflected in the hourly data gathered from sensors.

Ambient Temperature (AT): Ranges from 1.81°C to 37.11°C. Atmospheric Pressure (AP): Ranges from 992.89 to 1033.30 mbar. Relative Humidity (RH): Ranges from 25.56% to 100.16%. Vacuum (Exhaust Steam Pressure, V): Ranges from 25.36 to 81.56 cm Hg. Full Load Electrical Power Output (PE): Target variable ranges from 420.26 MW to 495.76 MW. Machine learning algorithms are used to establish a relationship between these independent variables (inputs) and the dependent variable (output). Each instance in the dataset, denoted as (Xi,Yi), represents a set of input-output pairs. The goal of machine learning regression methods is to learn a mapping function Y = f(X) that accurately predicts the electrical power output based on the input variables.

The regression technique aims to minimize the difference between the actual output (Y)of the system and the predicted output (Y) derived from the training dataset. This process involves finding the optimal function that captures the complex relationships between ambient conditions and turbine performance, as illustrated in Figure 1.

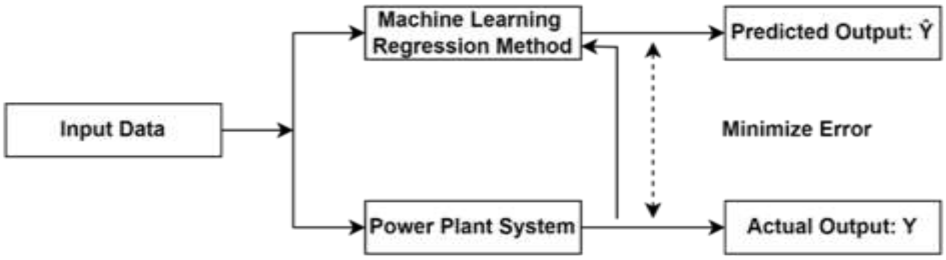

Figure 1-1 A machine learning regression method using real system data to predict.

3. MATLAB Regression Leaner

A graphical tool with an intuitive MATLAB interface for regression analysis is the MAT- LAB Regression Learner. Regression model training and performance evaluation are made easier with this application. It also streamlines the data preparation and import procedure. It is compatible with multiple regression techniques, including decision trees, neural networks, ensemble techniques, support vector machines, and linear regression. It has capabilities for choosing pertinent features and modifying hyperparameters to improve model accuracy. The objective is to forecast the electrical power output (PE), which is the dependent variable, us- ing the Regression Learner with our dataset. A five-fold cross-validation strategy is selected in order to guarantee robustness and avoid overfitting. Using this method, the data is divided into five subsets. The model is trained on four of the subsets, and it is validated on the fifth. This procedure is repeated for all possible combinations. To evaluate how effectively a model generalizes to unknown data, validation techniques such as holdout validation, resubstituting validation, and cross-validation are crucial. A model is said to be overfitting if it grows unduly complex, fits the training set too closely, and exhibits poor performance on fresh data. Before training models, it is essential to choose a suitable validation scheme in order to reliably assess their performance. The validation strategy selected affects every training model, guaranteeing an equitable and trustworthy evaluation.

The MATLAB Regression Learner provides a strong and effective means of investigating different regression models [10]. Accuracy is ensured by reliable validation methods like cross- validation, which also helps to mitigate problems like overfitting. Validation methods which are essential for assessing how well a model generalizes to unseen data.

3.1 Cross-Validation

Select what number of folds (or divisions) to use to divide the data collection. If k folds are selected, the app will,

i. The data is divided into k distinct sets or folds.

ii. For each fold in the validation:

•Trains a model using observations from the training fold (not observations from the vali- dation fold).

•Assesses model performance using validation-fold data

iii. Determines the mean validation error across all folds.

The predictive accuracy of the final model trained using the entire data set is well-estimated using this strategy. Although the method necessitates several fits, it effectively utilizes all the data, making it suitable for tiny data sets [7].

3.2 Holdout Validation

A portion of the data should be chosen to serve as a validation set. A model is trained on the training set using the programmed, and its performance is evaluated using the validation set. Holdout validation is only applicable for big data sets because the model used for validation is based only on a subset of the data. The complete data set is used to train the final model.

3.3 Resubstituting Validation

Overfitting is not protected against. Using all the data, the programmed trains determine the error rate using the same data. In the absence of any further validation data, you obtain an imprecise evaluation of the model's performance on new data. Stated otherwise, it is likely that the training sample accuracy will be excessively high and the anticipated accuracy would be lower.

The fiollowing Regression models showed promise with smaller RMSEs

3.3.1 Ensemble bagged tress

Ensemblen bagged tress is a potent machine learning technique that aims to increase the robustness and accuracy of models. It operates by employing bootstrap samples of the training data to create several models, most commonly decision trees. With bootstrap sampling, subsets of the training data are chosen at random using replacement. This implies that certain observa- tions may be repeated in a subset while others may not appear at all. A different decision tree is trained using each of these subsets. Bagging's main concept is to increase model diversity by having them trained on several data subsets. By lowering the model's variance, this variety can improve the model's ability to generalize results to previously untested data. A final prediction is produced by combining the forecasts of each individual decision tree once it has been trained. For regression problems, this aggregation procedure can entail averaging the predictions, and for classification tasks, voting. It is also possible to improve ensemble bagged trees by including other methods like boosting or random forests. For example, random forests add another layer of unpredictability to bagging by employing a random selection of features at each decision tree split in addition to sampling subsets of the data. Contrarily, boosting focuses on training models in a sequential manner where each new model fixes the mistakes of the previous one, eventually creating a powerful predictive model through repetitions. The bagging algorithm is exemplified by the following:

1. For every training data bootstrap sample i=1…, B

a. To construct a fresh bootstrap sample Xi, randomly select a portion of the training data with replacement.

b. Using a portion of the characteristics at each split, train a decision tree Ti on the bootstrap sample Xi.

c. Keep theTidecision tree on hand.

2. To anticipate a new input vector x, compute the ensemble's expected values for each decision tree, then average them to produce the final prediction:

$$\hat{Y}(x) = \frac{1}{B}\sum_{i=1}^{B} T_i(x) \qquad (1\text{-}1)$$

here Ti(x) is the decision tree, i is the predicted value for input vector x. The bagging algorithm is mathematically represented by this formula, where the final prediction is obtained by averaging the predictions of various decision trees. The goal of bagging is to reduce variance and enhance performance by building a diversified group of models that are each trained on slightly different subsets of the data. Ensemble bagged trees are effective because they harness the power of multiple models trained on varied data subsets, thereby improving predictive ac- curacy and resilience to overfitting compared to individual models trained on the entire dataset. This makes bagging particularly useful for creating robust regression models that can handle complex datasets and generalize well to new, unseen data.

3.3.2 Gaussian Process Regression

(GPR) is a probabilistic machine learning method used for regression tasks. Unlike tra- ditional regression methods that assume a parametric form for the relationship between inputs.

Exponential GPR, specifically using the exponential covariance function, is effective at captur- ing smooth patterns in data. It is particularly useful when the relationship between inputs and outputs exhibits gradual changes or correlations that can be modeled effectively by an expo- nential decay in similarity between data points as their distance increases. A specific kind of GPR called exponential GPR makes use of an exponential covariance function. In mathematical terms, the exponential covariance function is:

$$K(x - x') = \sum f^2 \times e^{\left(-0.5 \times \frac{\|x - x'\|^2}{l^2}\right)} \tag{1-2}$$

where $||x-x'||^2$ is the squared distance between the input vectors x and x' and f^2 is the variance of the function values. *l* is a length scale parameter. The function's smoothness is determined by the length scale parameter *l*. When *l* is large, the function changes slowly as the input variables change, whereas a small l causes the function to change more quickly. The GPR model esti- mates the posterior distribution over the function values at x using the training data to produce a prediction for a new input vector x. The training data and the covariance function determine the mean and variance of the posterior distribution, which is a Gaussian distribution. Exponen- tial GPR is a specific variant of GPR that uses an exponential covariance function, which can capture smooth patterns in the data.

3.3.3 Support Vector Machine

SVM is an effective machine-learning technique used for both regression and classification tasks [12]. Finding a hyperplane that divides the data into two classes and maximizes the distance between the hyperplane and the nearest data points is the aim of SVM. By predicting the value of a continuous output variable rather than a binary class label, SVM can also be utilized for regression applications. An SVM version known as a Gaussian SVM sometimes referred to as a Radial Basis Function (RBF) SVM, uses a Gaussian kernel to map the input data into a high-dimensional space. A definition of the Gaussian kernel is:

$$K(x_i - x_j) = e^{-\gamma \|x_i - x_j\|^2} \tag{1-3}$$

where is the squared distance between the input vectors, gamma is a hyperparameter that regulates the kernel's width, and *xi* and *xj* are the input vectors. The radial form of the Gaussian kernel constricts as one moves away from the center. In order to improve the performance nof a Gaussian SVM on a particular dataset, the hyperparameters of the model are frequently adjusted. The most crucial Gaussian SVM hyperparameters are C, which regulates the trade-off between the margin and classification error, and gamma, which regulates the Gaussian kernel's width. The best values for these hyperparameters are frequently found via grid searches or other optimization methods. With an RMSE value of 3.51, the ensemble bagged tree was the best of the three regression models, followed by Gaussian Process Regression (Exponential GPR), Gaussian Support Vector Machine (Fine Gaussian SVM), and Gaussian Process Regression. These are the top three models, even though there are alternative models with various RMSEs.

4.Results

4.1 DATASET DESCRIPTION

The dataset used in this study comprises four input variables and one target variable col- lected over a period of six years (2006-2011). It consists of 9,568 data points gathered from a combined cycle power plant (CCPP) operating at full capacity across 674 distinct days. Input Variables: The input variables represent typical hourly measurements obtained from sensors at the plant. These variables include: Ambient Temperature (T) Ambient Pressure (AP) Rela- tive Humidity (RH) Exhaust Vacuum (V) Target Variable: The target variable is the full-load electrical power output (PE), which measures the average hourly output when the power plant operates at base load. This data is crucial for assessing the plant's operational efficiency and performance.

4.2 Data Preprocessing

The collection initially included some erratic and noisy data, mostly as a result of electrical disruptions influencing sensor readings. Furthermore, during preprocessing, data points with the power plant operating at less than 420.26 MW were deemed incompatible and removed. At first, there were 674 distinct daily files in the dataset, all in the.xls format. These files were cleaned and merged into a single dataset to guarantee data consistency and integrity. After removing duplicate entries, a consolidated dataset that was prepared for analysis was obtained. Additional preparation procedures, described in [13], included randomly rearranging the dataset in order to reduce bias and improve the stability of ensuing analyses. To make the final dataset easier to use and more accessible in analytical tools, it was transformed to the.xlsx format.

5. PREDICTION ACCURACY

The prediction accuracy of each machine-learning regression technique is used to evaluate the overall agreement between the expected and real values. The performance metrics employed in this study to evaluate the prediction accuracy are the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) for continuous variables.

iv. Mean Absolute Error (MAE): Mean absolute error is the average of all test cases' antic- ipated and actual values, without taking direction into account [40].

$$\text{MAE} = \frac{(|a_1 - c_1| + |a_2 - c_2| + \cdots + |a_n - c_n|)}{N} \quad 1\text{-}4)$$

v. Root Mean Square Error (RMSE): is a commonly used indicator of discrepancies be- tween values predicted by a model or estimator and the values obtained from the process being modelled or estimated [9.RMSE =

$$\text{RMSE} = \sqrt{\frac{(a_1 - c_1)^2 + (a_2 - c_2)^2 + \cdots + (a_n - c_n)^2}{n}} \quad (1\text{-}5)$$

In the two equations above, a represents the output's actual value and c its expected value. A lower value indicates a more accurate model in all the error assessments, with a value of 0 indicating a statistically flawless model [11].

Table 1-1 Training Results of Ensemble Bagged Trees

Metric	**Value**
RMSE (Validation)	3.51
R-Squared (Validation)	0.96
MSE (Validation)	12.32
MAE (Validation)	2.5484
Prediction Speed	9200
Training Time	559.92 sec
Model Size	6 MB

Table 1-2 Ensemble Hyperparameters Used in Model Training

Preset	Bagged Trees
Minimum leaf size	8
Number of learners	30
Number of predictors to sample	Select all

5.1 Ensemble Bagged Trees Results (Summary)

Partial dependence charts (PDPs) are a useful tool for visualizing the projected response of a trained regression model. PDPs display the marginal impacts of each predictor. Partial de- pendence graphs illustrate how the expected values of the output variable fluctuate in response to changes in the value of a single input variable, with other variables being maintained con- stant. The partial dependence graphs display the fitted model along with the output variable's projected values at different input variable values. In addition to identifying any nonlinearities or interactions between the input variables, these plots are used to assess the correlation between the input and output variables.

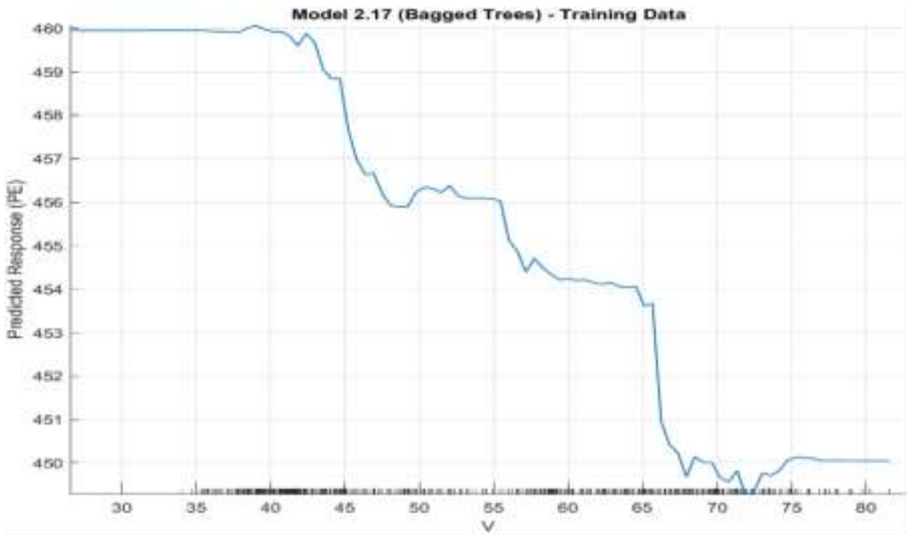

Figure 1-2 Partial Dependence Plot PE vs V

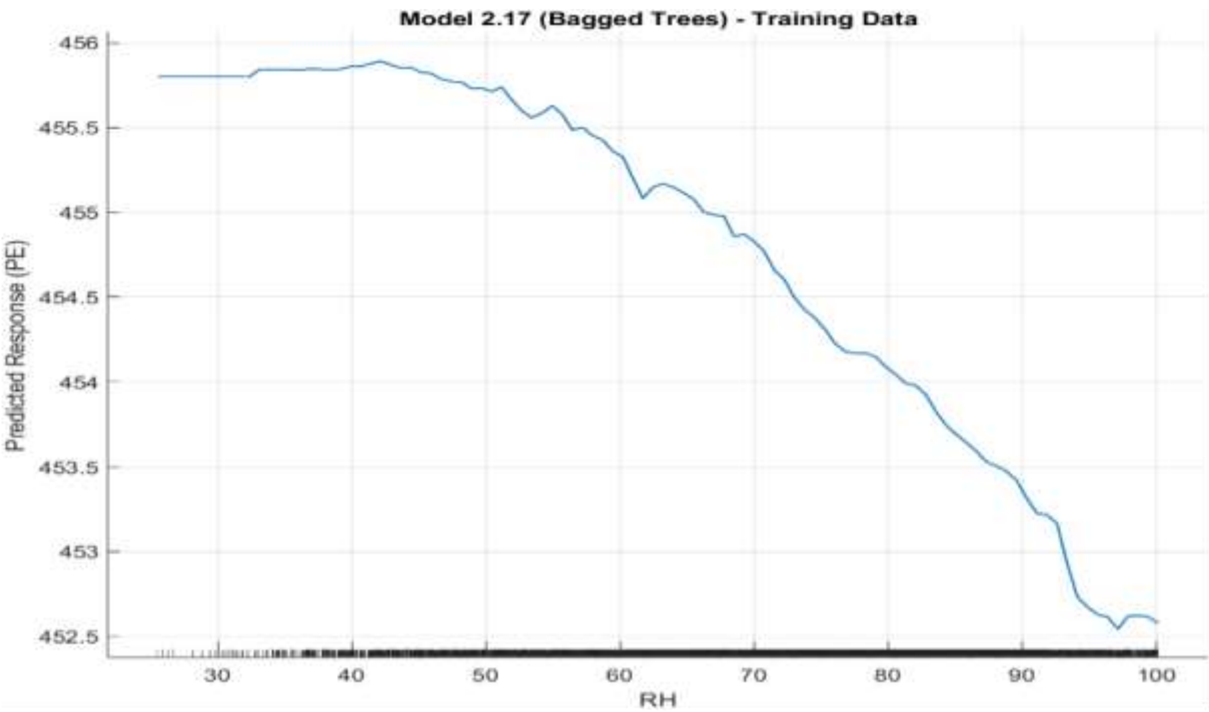

Figure 1-3 Partial Dependence Plot PE vs RH

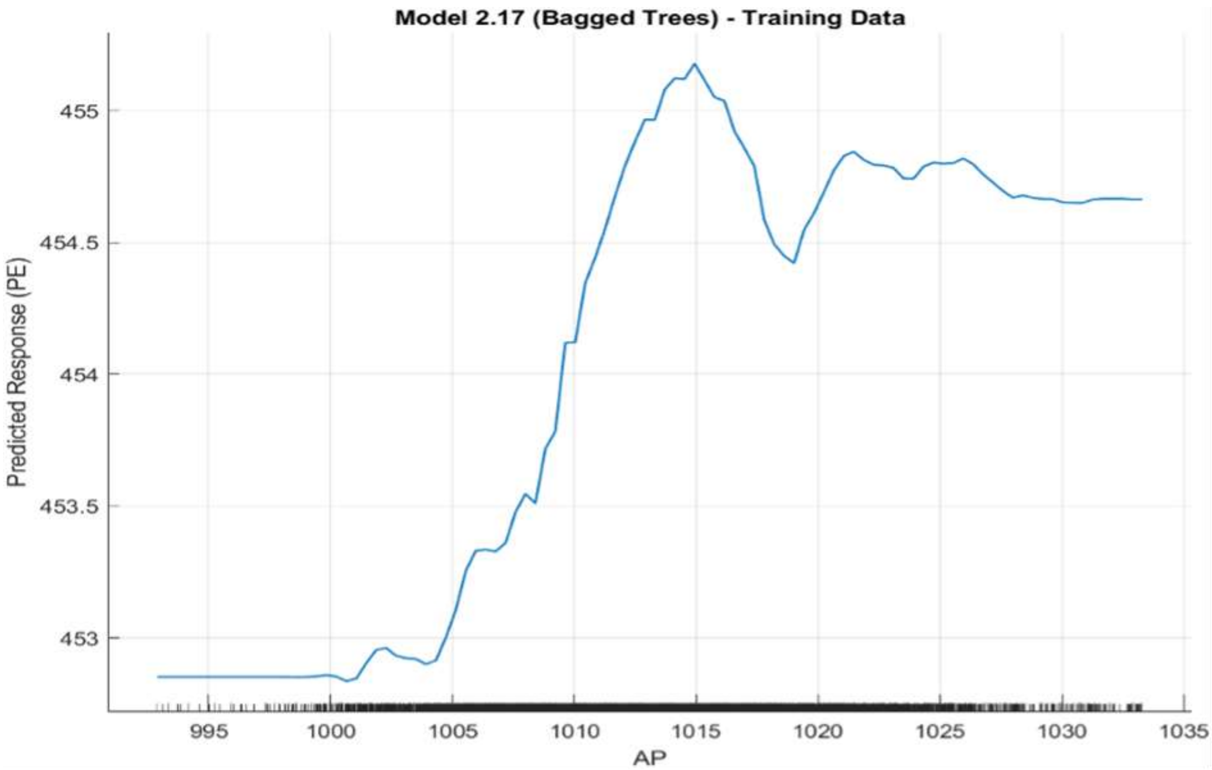

Figure 1-4 Partial Dependence Plot PE vs AP

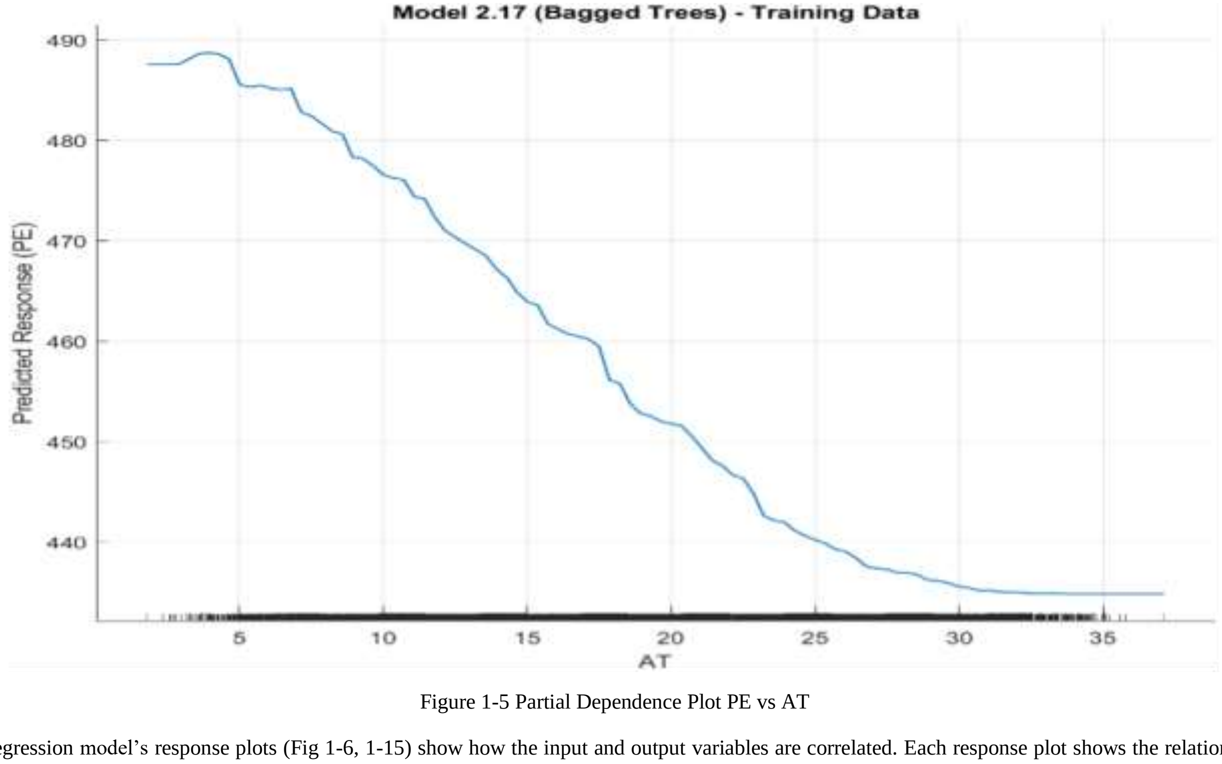

Figure 1-5 Partial Dependence Plot PE vs AT

A regression model's response plots (Fig 1-6, 1-15) show how the input and output variables are correlated. Each response plot shows the relationship between a specific input variable and the output variable when all other input variables are held constant. The response plots display both the fitted model and the actual data points. These graphs assess the degree to which the input and output variables have a linear or nonlinear relationship, as well as how well the model matches the data.

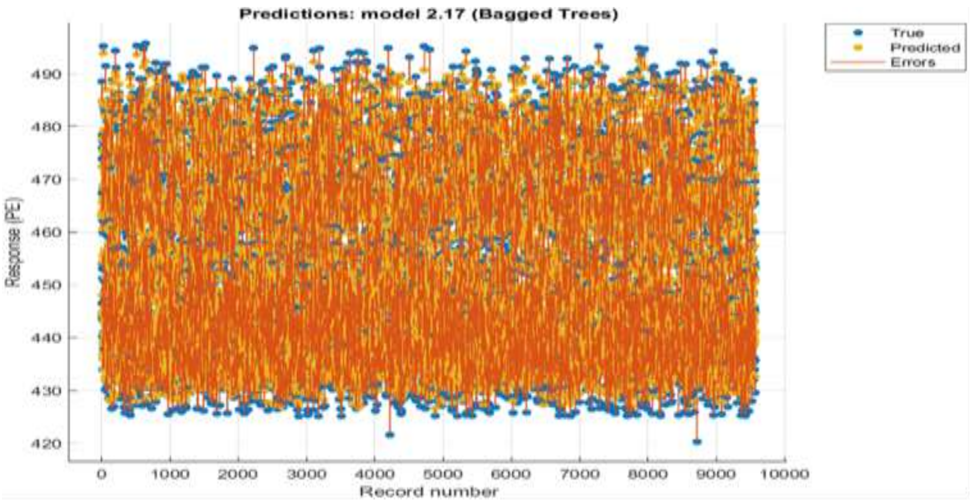

Figure 1-6 Response Plot PE vs Record Number with errors

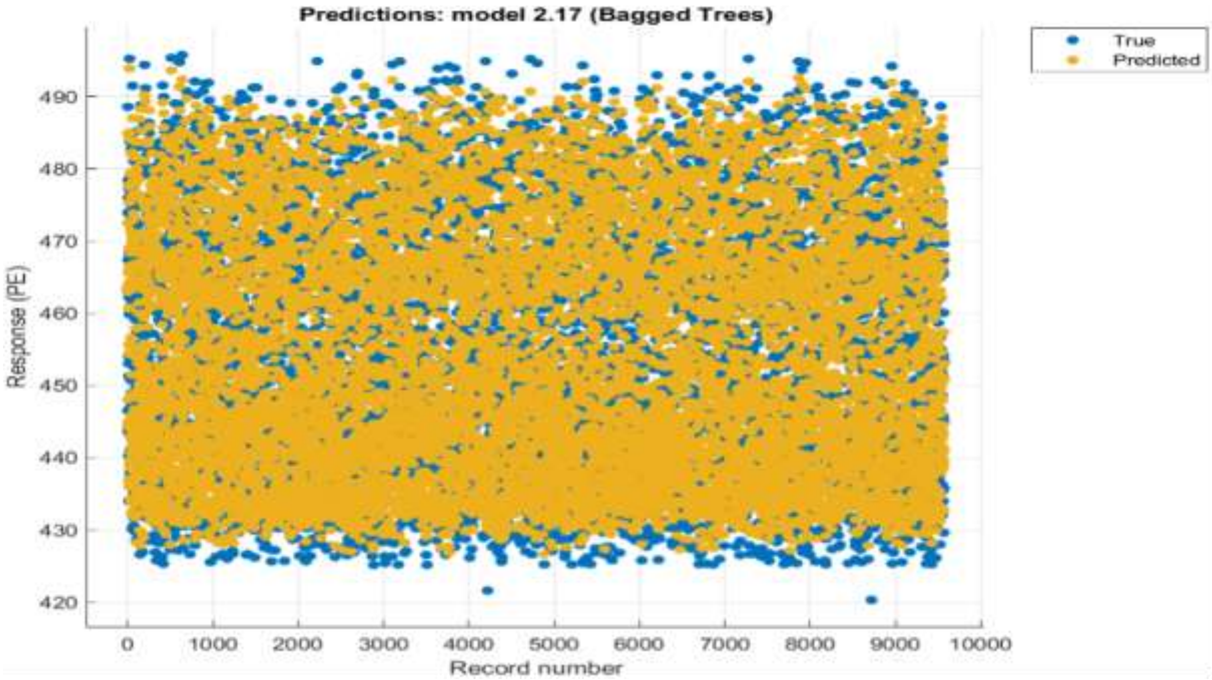

Figure 1-7 Response Plot PE vs Record Number

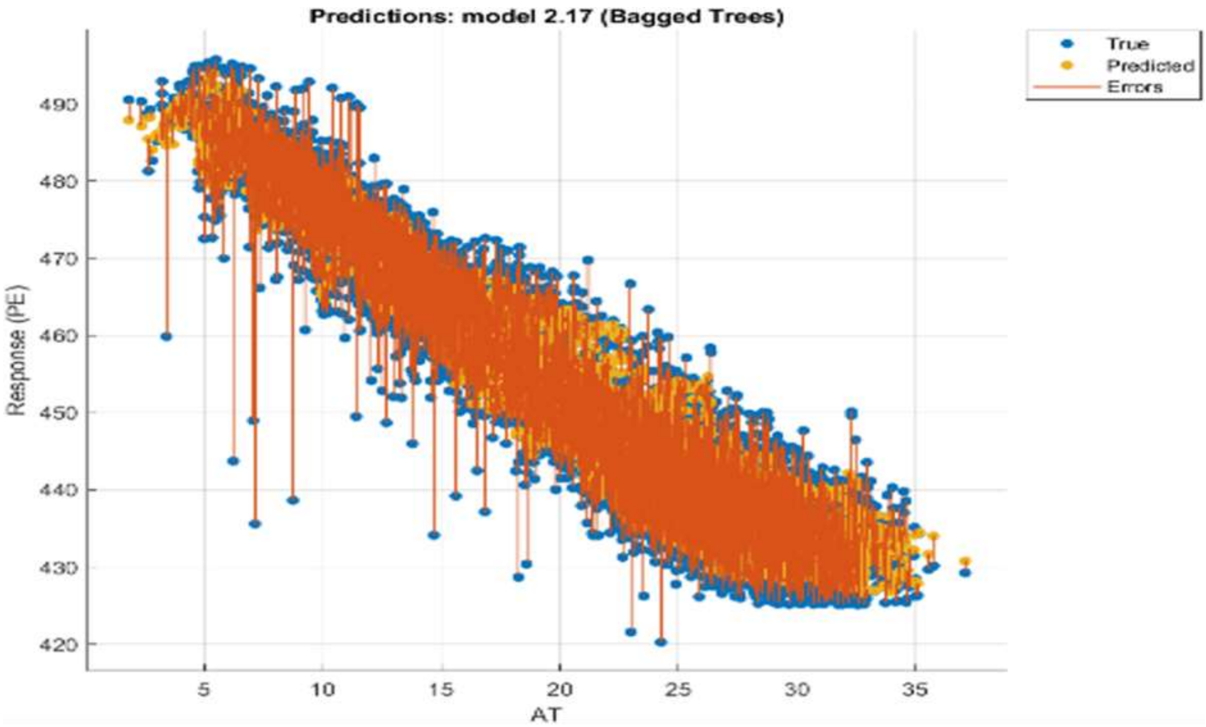

Figure 1-8 Response Plot PE vs AT with errors

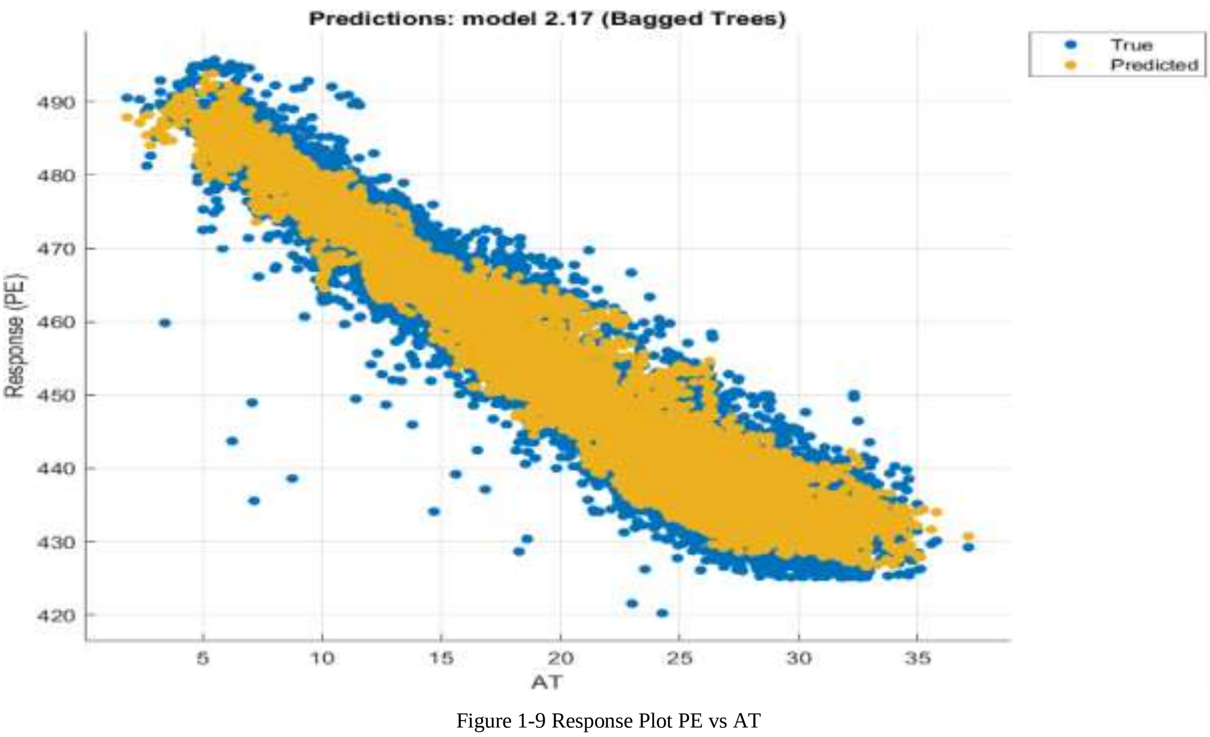

Figure 1-9 Response Plot PE vs AT

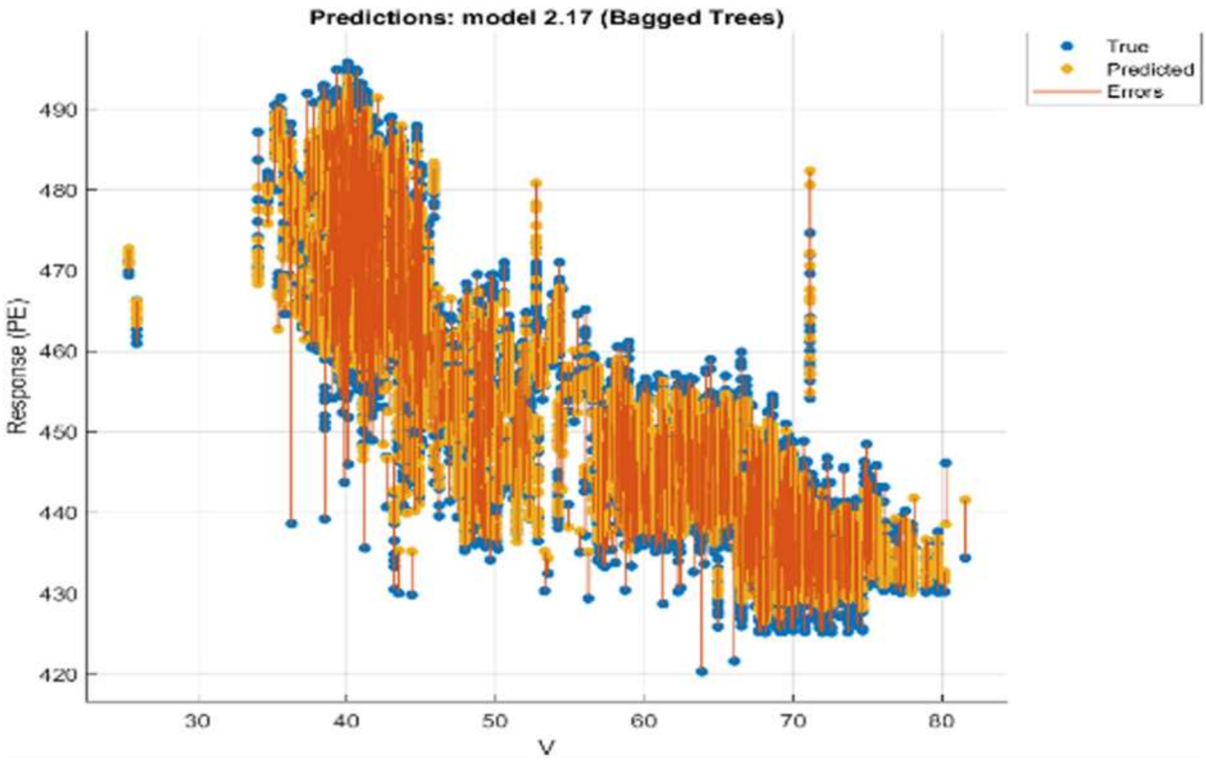

Figure 1-10 Response Plot PE vs V with errors

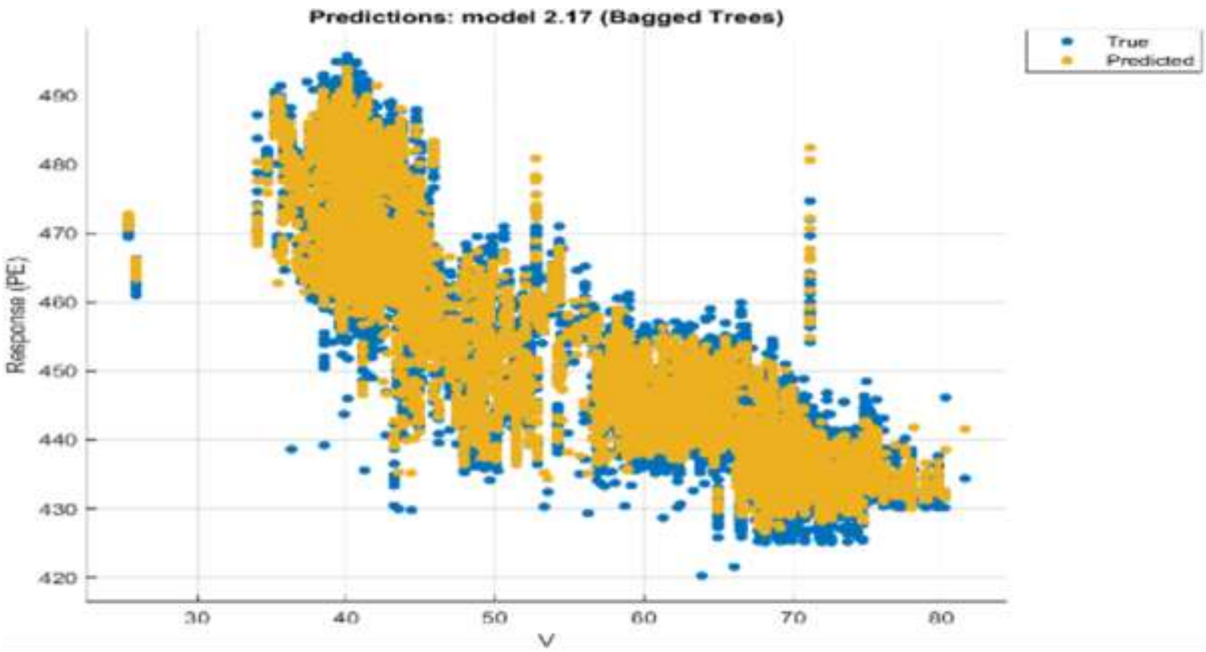

Figure 1-11 Response Plot PE vs V

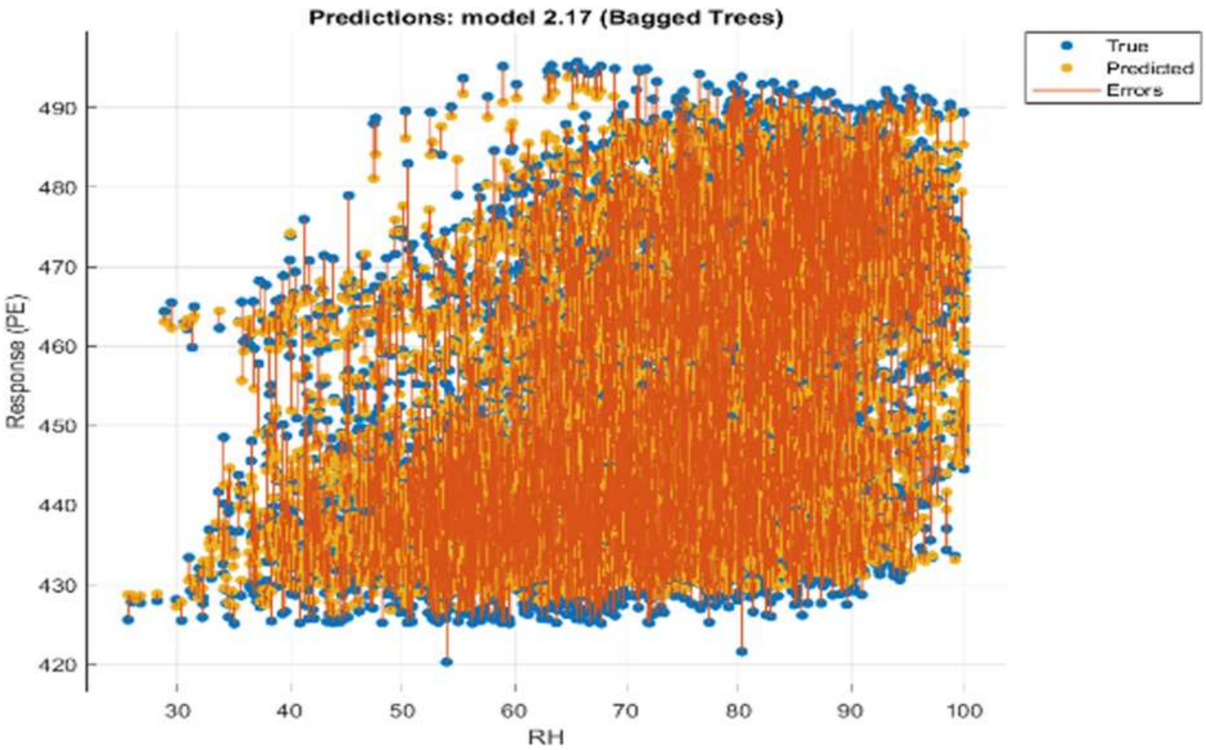

Figure 1-12 Response Plot PE vs RH with errors

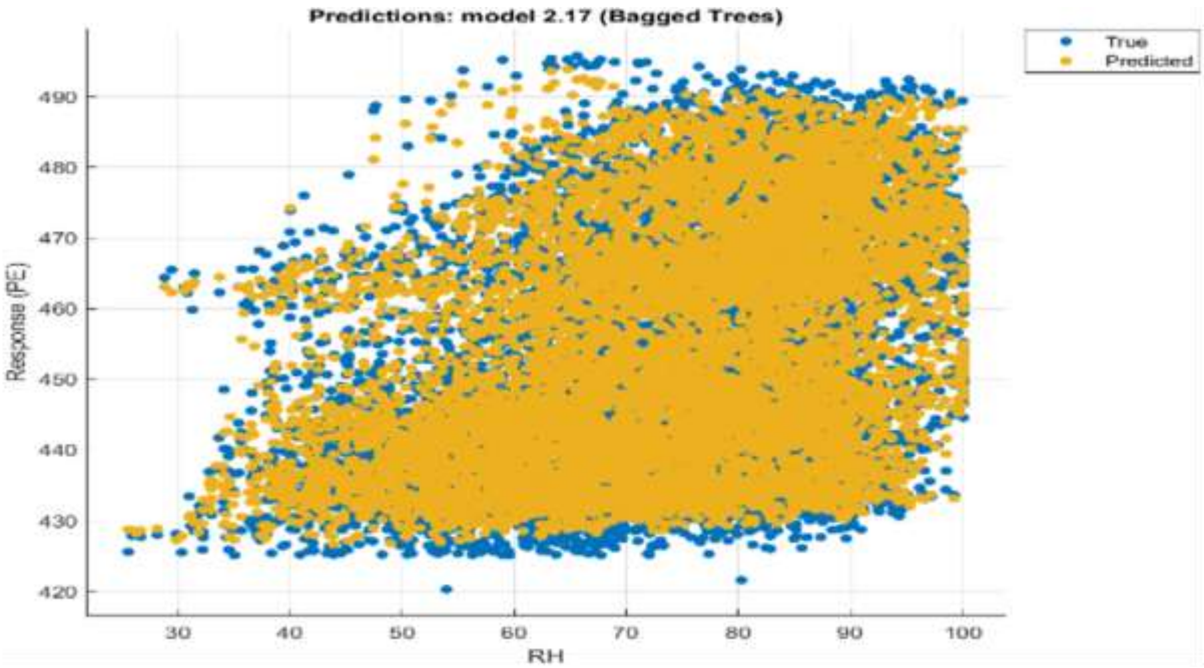

Figure 1-13 Response Plot PE vs RH

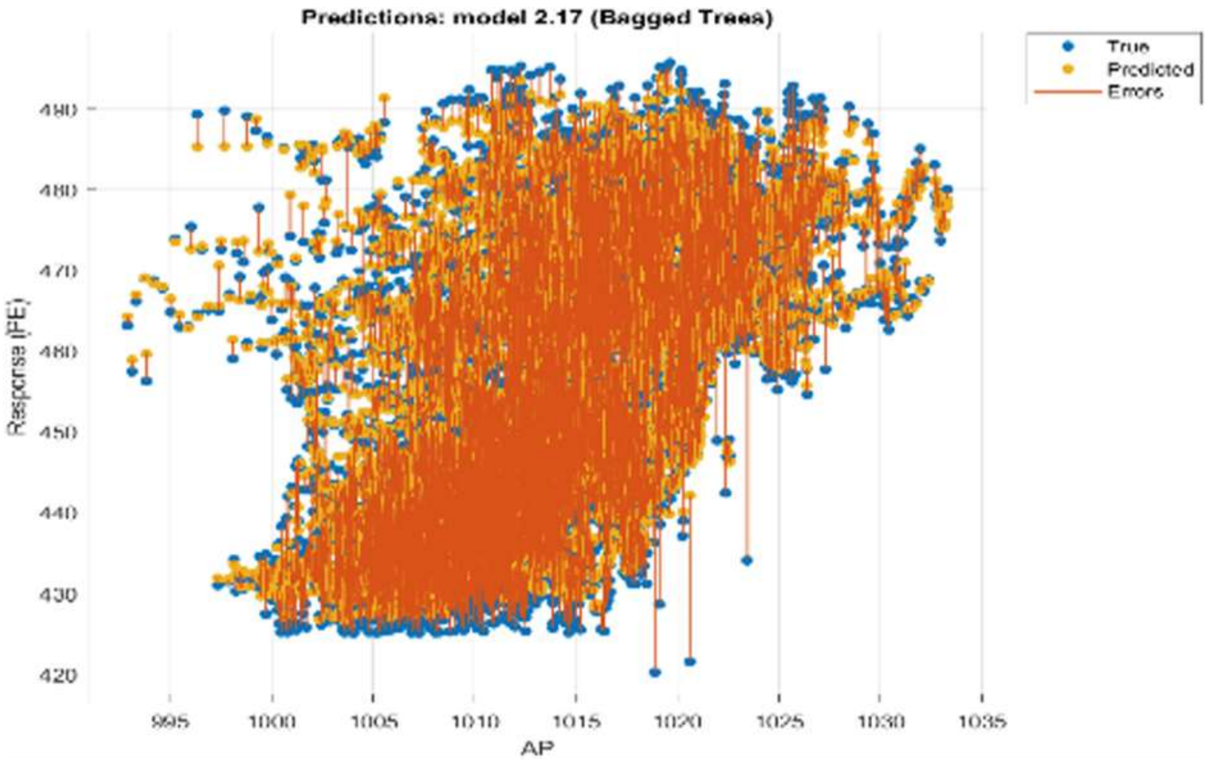

Figure 1-14 Response Plot PE vs AP with errors

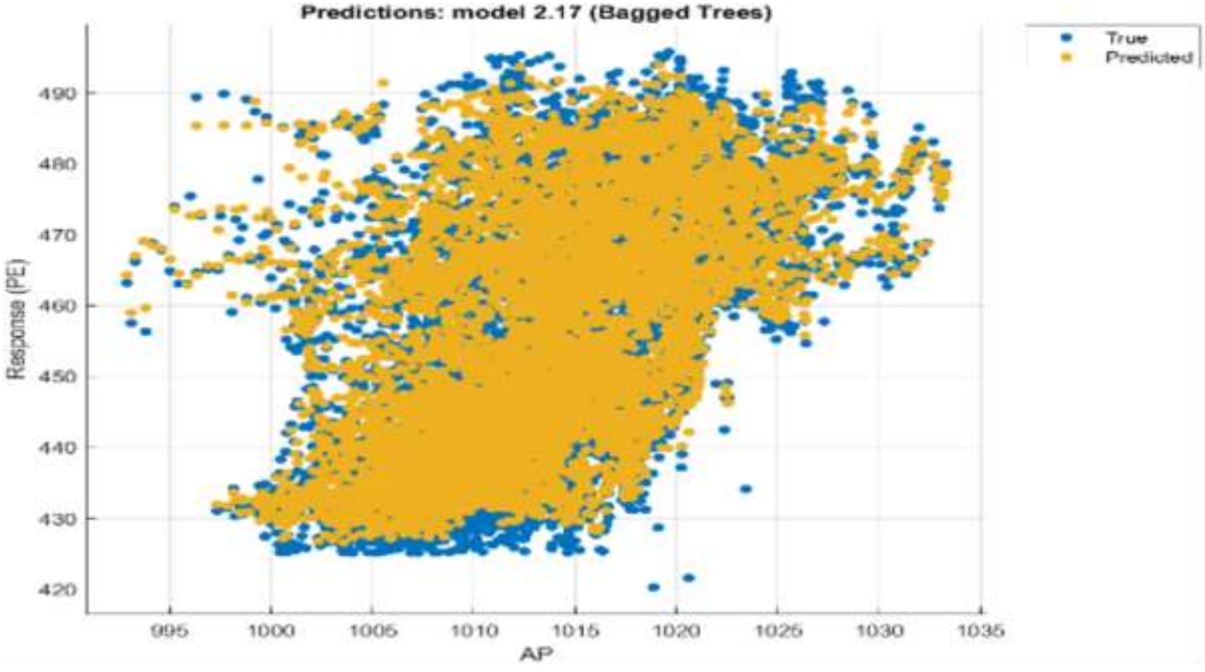

Figure 1-15 Response Plot PE vs AP

The performance of the regression model and how well it forecasts different response values are assessed by the expected vs. real plot. Because the projected response of a perfect regression model equals the true response, all the points lie on a diagonal line. The vertical distance from the line to any given location equals the forecast error for that point. A well-designed model produces predictions that are distributed throughout the line with small errors.

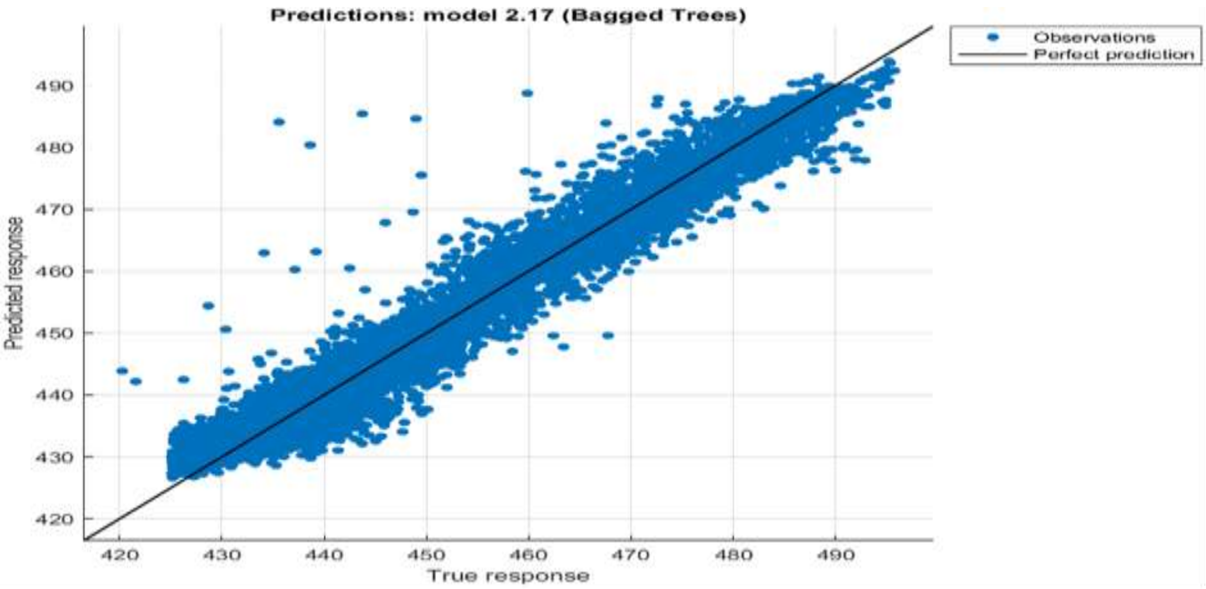

Figure 1-16 Validation Predicted vs. Actual Plot (True response

The residuals plot (Figrs 1-17, 1-23) assesses how well the model predicts the outcome variable by displaying the difference between observed values (actual responses) and predicted values. When these residuals are symmetrically distributed around zero, it indicates that the model accurately predicts responses across various data ranges. This symmetry supports the reliability and robustness of the model's predictions, validating its effectiveness in capturing the underlying patterns in the data.

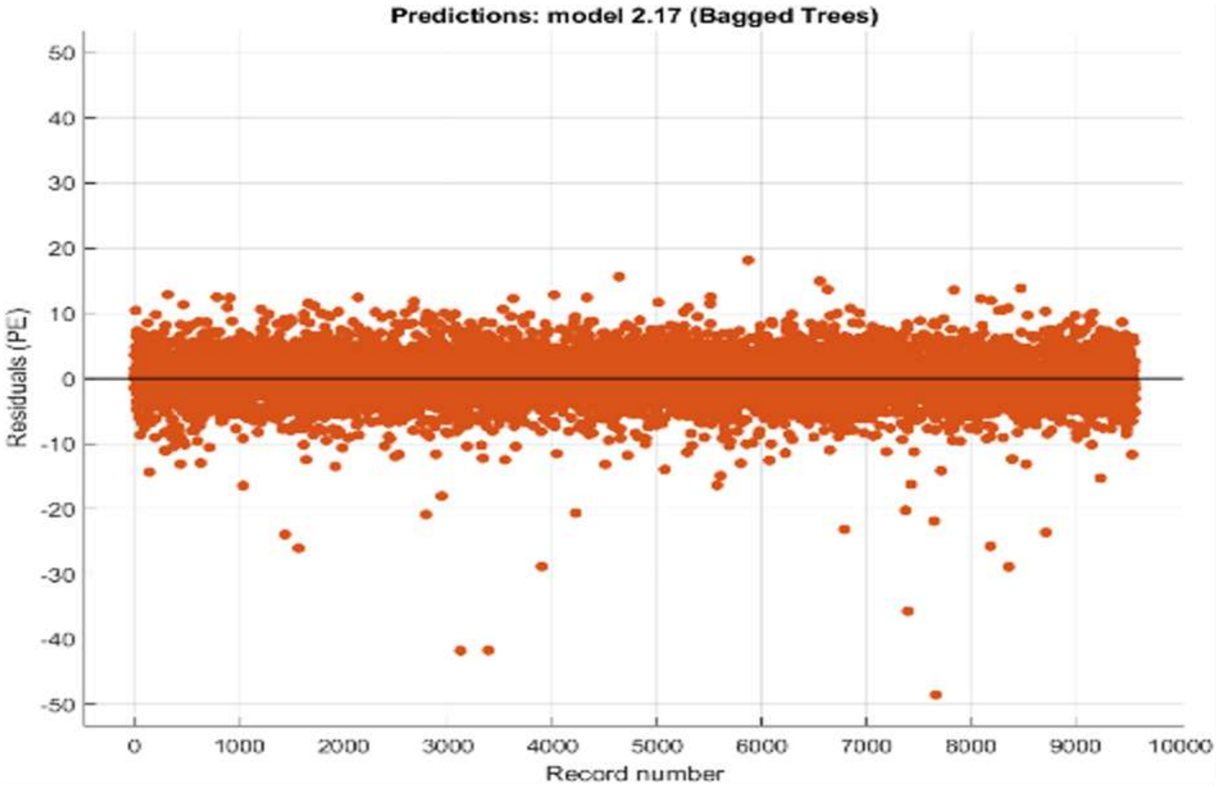

Figure 1-17 Validation Residuals Plot Record number

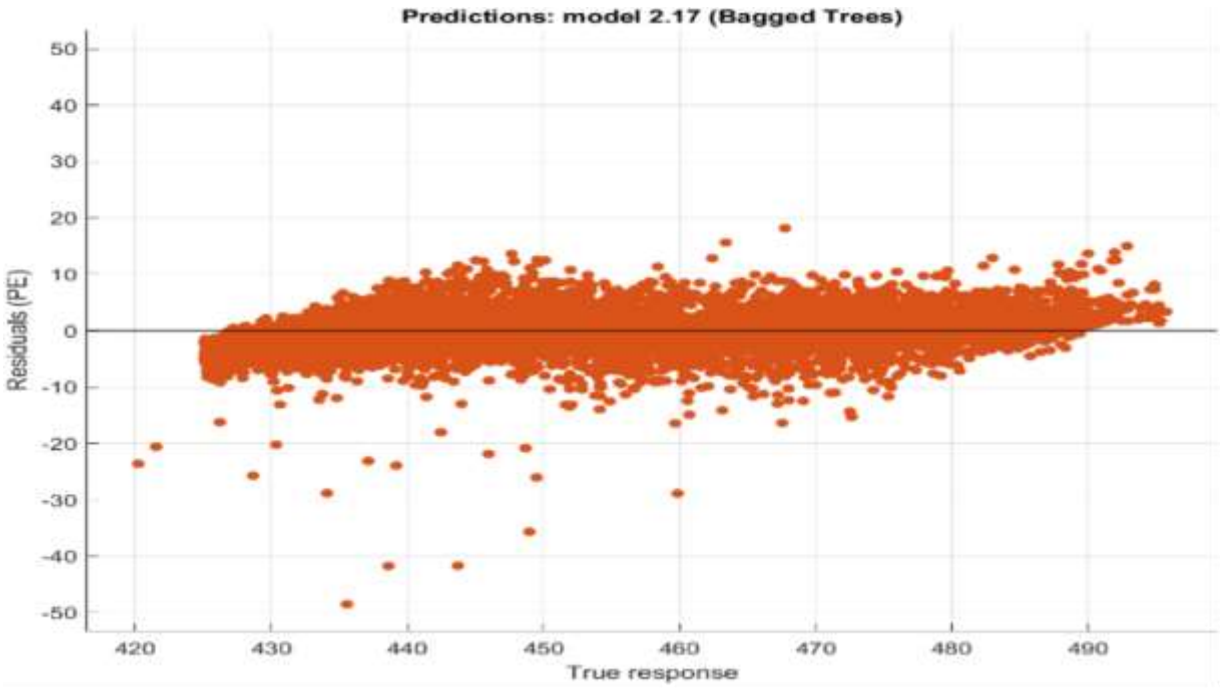

Figure 1-18 Validation Residuals Plot True response

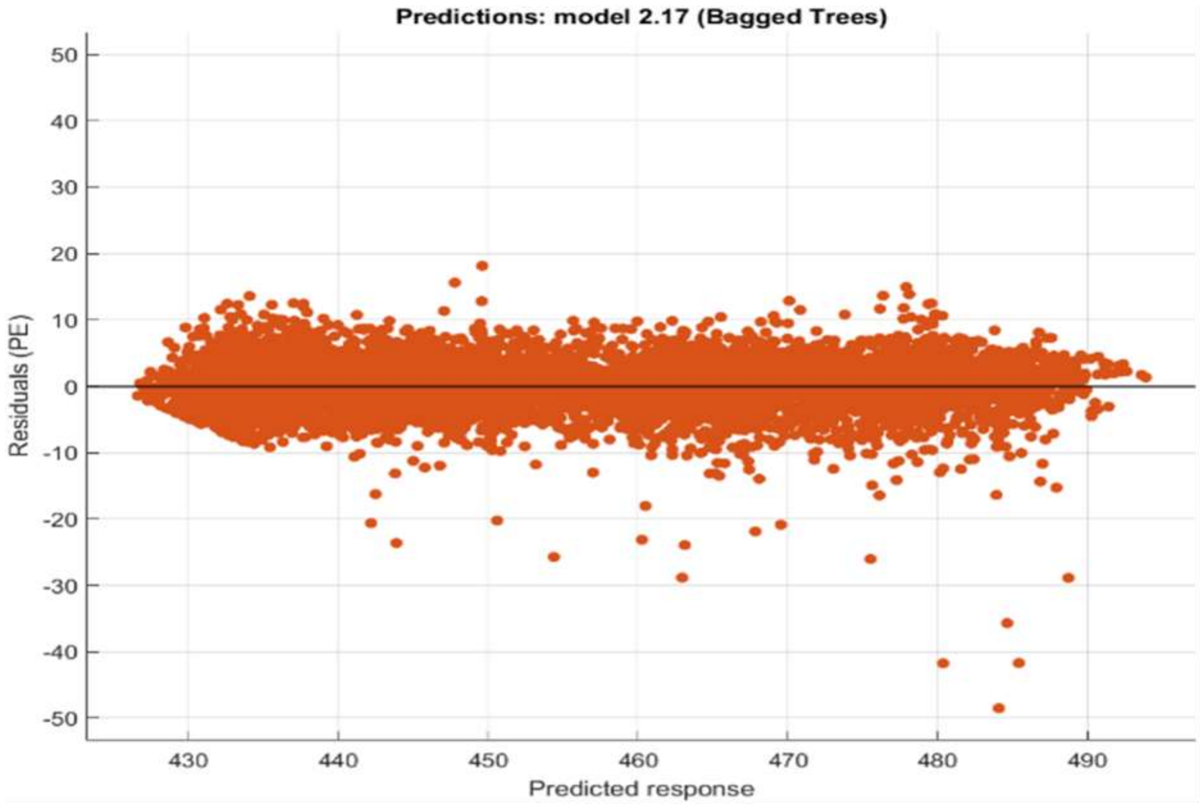

Figure 1-19 Validation Residuals Plot Predicted response

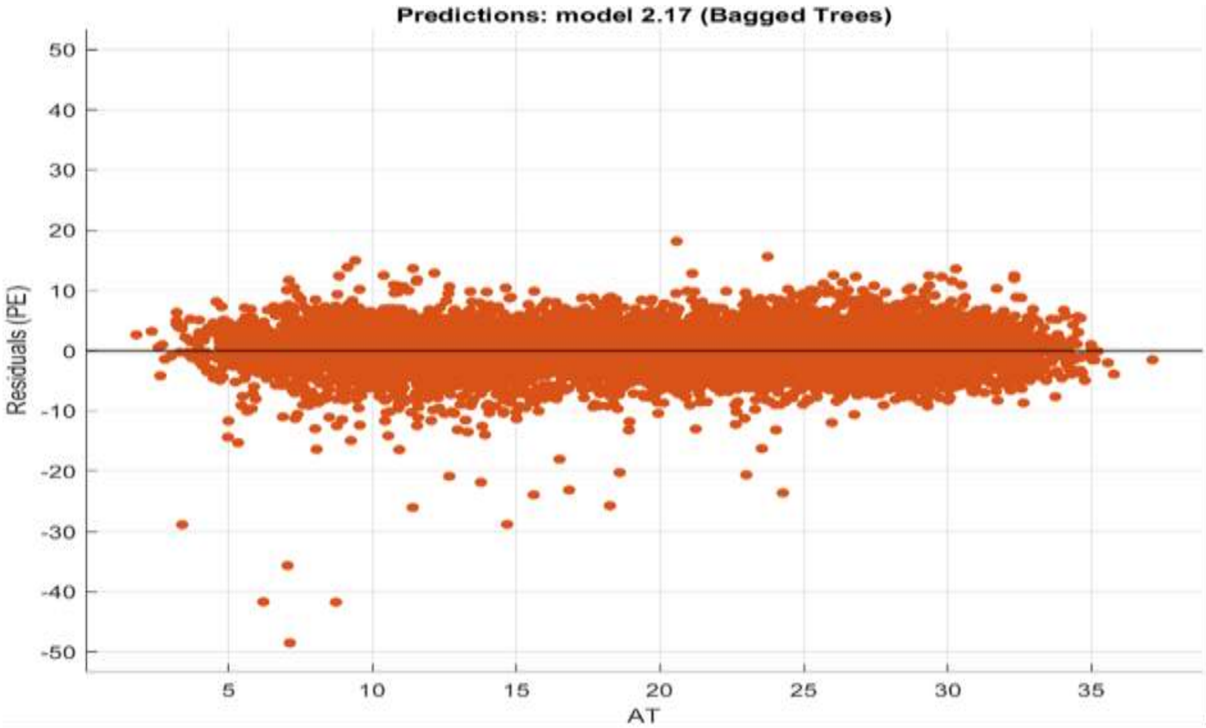

Figure 1-20 Validation Residuals Plot Predictor AT

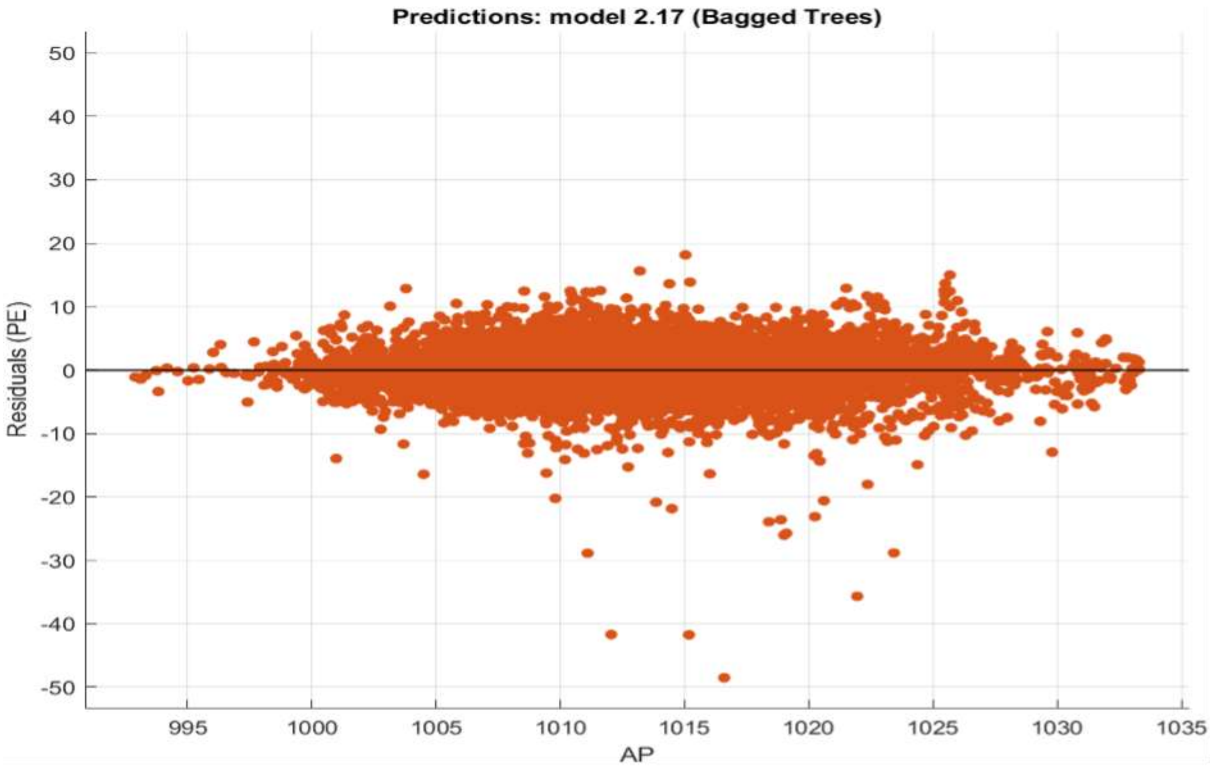

Figure 1-21 Validation Residuals Plot Predictor AP

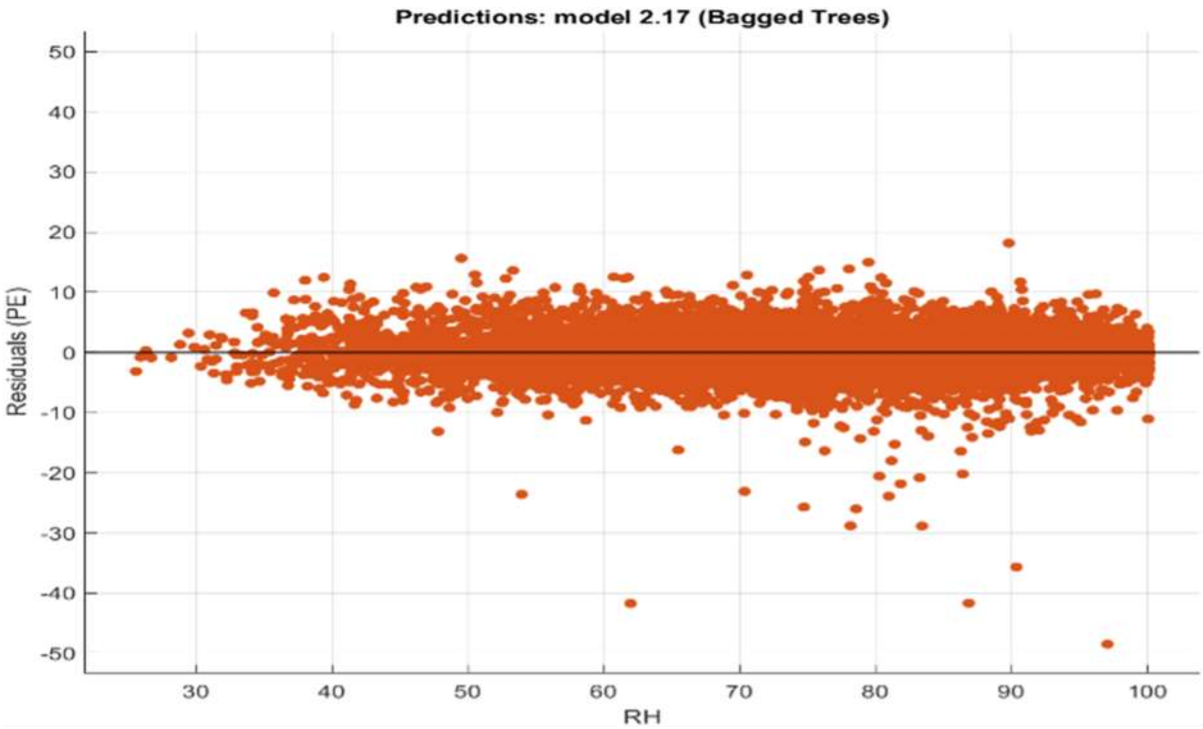

Figure 1-22 Validation Residuals Plot Predictor RH

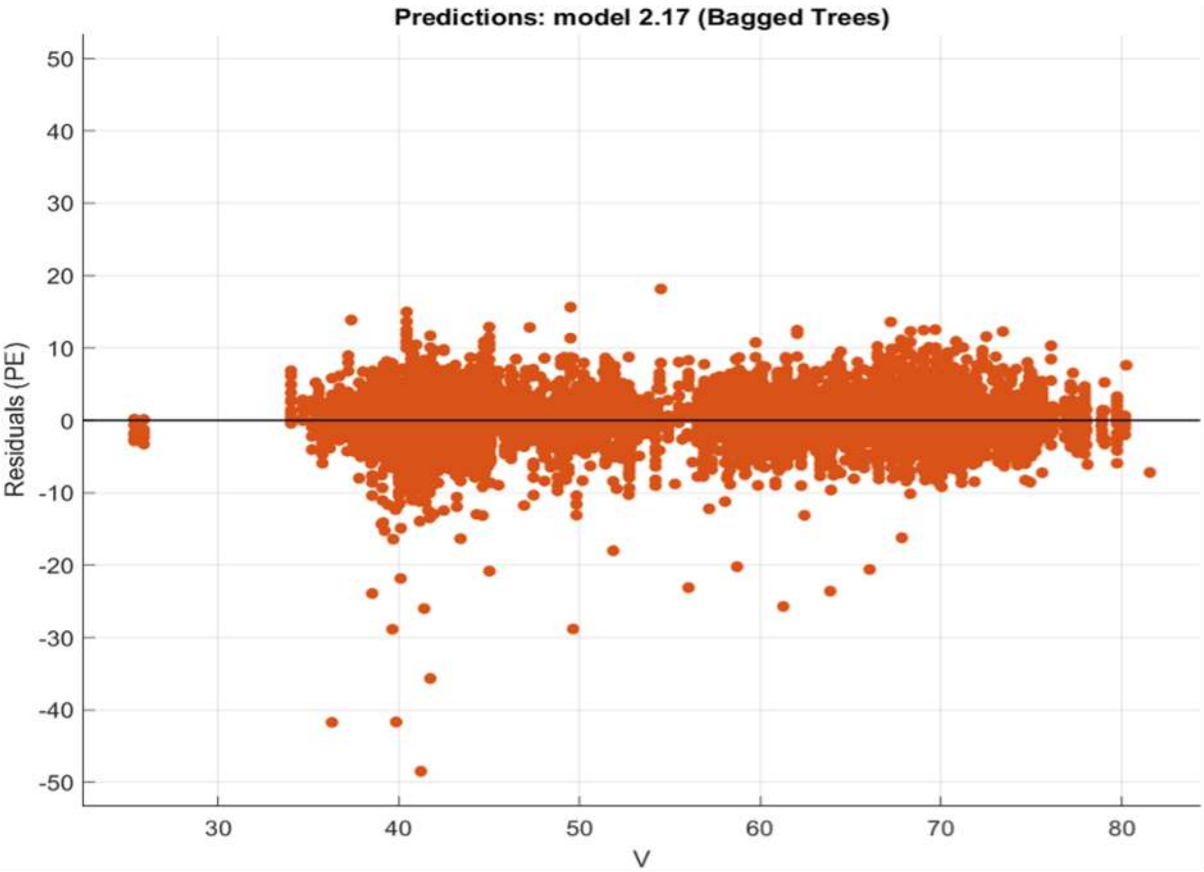

Figure 1-23 Validation Residuals Plot Predictor V

In terms of prediction accuracy, the Ensemble Bagged Trees model fared better than all other trained models based on the evaluation criteria of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Regression analysis often uses RMSE and MAE as metrics to evaluate how well a model predicts the outcomes in comparison to the actual values. When the RMSE is smaller, it means that, on average, the model's predictions are more accurate than the actual values across the dataset. In a similar vein, a lower MAE denotes fewer absolute discrepancies between the model's predictions and the actual values. Out of all the trained models in this investigation, the Ensemble Bagged Trees model has the lowest RMSE (3.51) and MAE (2.5484).

These findings suggest that, for the study's purposes, the Ensemble Bagged Trees model offers the most precise and trustworthy estimates of the full-load electrical power output (PE). Compared to the other regression models examined, the model's better performance in mini- mizing both RMSE and MAE shows it can estimate the power output more precisely. Because of this, it is a better option for real-world scenarios where precise power generation forecasting is essential. Its robustness and generalizability are highlighted by the test dataset's validation of its effectiveness, which further supports its appropriateness as a predictive tool for yet-to-be- seen data in related circumstances. Thus, the Ensemble Bagged Trees model is the best option for forecasting the results of full-load electrical power production (PE) in this study setting, according to these evaluation criteria.

Table 3. summarizes the performance of each of the models based on their RMSE and MAE values. The RMSE values range from 3.5 to 19, while the MAE values range from 12 to

20. Based on the evaluation criteria of RMSE and MAE, the best-performing model among the trained models was the Ensemble bagged Trees model. This model achieved the lowest RMSE and MAE values compared to the other models. This indicates that the Ensemble bagged Trees model is the most accurate and reliable model for predicting the outcomes of the PE.

Session: Regression Learning Models
Training Data: Folds5x2pp Observations: 9568
Predictors: 4 Predictor Names: AT, V, AP, RH
Response Name: PE
Validation: 5-fold cross-validation

Favorite	Model Nu	Model Typ	RMSE (Val	MSE (Valic	RSquared (	MAE (Valic
0	1	Tree	4.057748	16.46532	0.943478	2.895093
0	2.1	Linear Reg	4.559213	20.78642	0.928644	3.626664
0	2.2	Linear Reg	4.313089	18.60274	0.93614	3.39572
0	2.3	Linear Reg	4.562369	20.81521	0.928545	3.618798
0	2.4	Stepwise L	4.314021	18.61078	0.936113	3.396624
0	2.5	Tree	4.057748	16.46532	0.943478	2.895093
0	2.6	Tree	3.988862	15.91102	0.94538	2.922194
0	2.7	Tree	4.070934	16.5725	0.94311	3.074911
0	2.8	SVM	4.572124	20.90432	0.928239	3.6157
0	2.9	SVM	4.274009	18.26716	0.937292	3.344971
0	2.1	SVM	4.17735	17.45025	0.940096	3.2249
0	2.11	SVM	3.788197	14.35044	0.950738	2.74736
0	2.12	SVM	3.997725	15.9818	0.945137	3.022311
0	2.13	SVM	4.223329	17.83651	0.938771	3.285795
0	2.14	Ensemble	19.70124	388.1387	-0.33241	19.29093
1	2.15	Ensemble	3.52086	12.39645	0.957445	2.560782
0	2.16	Gaussian F	4.119533	16.97055	0.941743	3.189455
0	2.17	Gaussian F	3.948384	15.58973	0.946483	3.003631
0	2.18	Gaussian P	3.602984	12.9815	0.955437	2.638566
0	2.19	Gaussian F	3.896421	15.18209	0.947883	2.943818
0	2.2	Neural Ne	4.257516	18.12644	0.937775	3.34482
0	2.21	Neural Ne	4.21805	17.79194	0.938924	3.303175
0	2.22	Neural Ne	4.105241	16.853	0.942147	3.153918
0	2.23	Neural Ne	4.154957	17.26367	0.940737	3.230488
0	2.24	Neural Ne	4.151688	17.23651	0.94083	3.212725
0	2.25	Kernel	4.400511	19.3645	0.933525	3.367725
0	2.26	Kernel	4.216155	17.77596	0.938978	3.248573

Figure 1-24 Table 3 shows the Regression Models that were trained

In Table 4, the test results of the regression ensemble bagged tree model are presented. The model was applied to a dataset, and the predicted values for PE are shown in green next to the actual values. Notably, the difference between the actual PE values and the predicted PE values is extremely small, almost negligible. This indicates that the trained ensemble bagged trees model is highly accurate in estimating the output of electrical combined cycle power plants. Therefore, based on the results displayed in Table 4, it is evident that the ensemble bagged trees model stands out as the most precise method for predicting PE outcomes.

AT	V	AP	RH	PREDICTED PE	ACTUAL PE
8.34	40.77	1010.84	90.01	480.7472955	480.48
23.64	58.49	1011.4	74.2	446.0939461	445.75
29.74	56.9	1007.15	41.91	437.7298868	438.76
19.07	49.69	1007.22	76.79	453.4289784	453.09
11.8	40.66	1017.13	97.2	467.1673663	464.43
13.97	39.16	1016.05	84.6	470.4088598	470.96
22.1	71.29	1008.2	75.38	442.0906147	442.35
14.47	41.76	1021.98	78.41	464.0447311	464
31.25	69.51	1010.25	36.83	430.4586792	428.77
6.77	38.18	1017.8	81.13	484.7156225	484.31
28.28	68.67	1006.36	69.9	435.6998863	435.29
22.99	46.93	1014.15	49.42	451.8746789	451.41
29.3	70.04	1010.95	61.23	433.2994408	426.25
8.14	37.49	1009.04	80.33	480.5971884	480.66
16.92	44.6	1017.34	58.75	462.4331665	460.17
22.72	64.15	1021.14	60.34	450.3904488	453.13
18.14	43.56	1012.83	47.1	462.0726496	461.71
11.49	44.63	1020.44	86.04	472.369581	471.08
9.94	40.46	1018.9	68.51	472.9345233	473.74
23.54	41.1	1002.05	38.05	448.6517695	448.56
14.9	52.05	1015.11	77.33	463.7877481	464.82
33.8	64.96	1004.88	49.37	430.0345257	427.28
25.37	68.31	1011.12	70.99	438.3896744	441.76
7.29	41.04	1024.06	89.19	478.7089175	474.71
13.55	40.71	1019.13	75.44	470.2499789	467.21
6.39	35.57	1025.53	77.23	486.5845939	487.69
26.64	62.44	1011.81	72.46	439.2236045	438.67
7.84	41.39	1018.21	91.92	480.6857109	485.66
21.82	58.66	1011.71	64.37	448.836582	452.16

Figure 1-25 Table 4 shows some of the predicted output of the best Model

6.Discussions

The development of reliable methods for the correct prediction of yet-to-be-seen data is the main objective of machine learning research. Regression has shown to be a useful tool for forecasting in this study, as indicated by the encouraging outcomes that are covered in the Results section. In order to forecast the output of a combined cycle power plant (CCPP), which consists of two gas turbines, a steam turbine, and two heating systems, the research investigated several machine learning regression models. The goal of the study was to identify the precise variable or variables combined that had the biggest impact on the generation of full-load electrical power. In order to do this, over 25 regression models must be tested using 15 distinct combinations of the four main variables, AT, V, AP, and RH. It evaluated the best regression technique for forecasting the electrical power output under full load scenarios for various dataset subsets.

7.Conclusion

The study presented an alternative approach to predict the electrical power output of a combined cycle power plant (CCPP) operating at full load, opting for machine learning methods over traditional thermodynamical approaches. These methods were chosen for their ability to provide accurate predictions without the computational burden and potential unreliability of thermodynamic models that rely on numerous assumptions and nonlinear equations. Two main objectives guided the study: identifying the most influential variables in predicting power output and determining the most effective machine learning regression method for this prediction. The analysis involved testing 15 different combinations of four variables (AT, V, AP, RH) across 15 machine learning regression techniques. The results indicated that the subset containing all four parameters provided the most accurate predictions, achieving a Mean Absolute Error (MAE) of 2.5484 and Root Mean Square Error (RMSE) of 3.51. Specifically, the Bagging method with REP Tree predictor yielded the highest accuracy among the methods tested, with an MAE of

3.220 and RMSE of 4.239 on average. The developed predictive model has been implemented by the CCPP for forecasting hourly energy output using next day's temperature forecasts from the state's meteorology institute. Future research aims to enhance the model by improving the precision of ambient variable predictions and extending its application to different types of power plants. This approach not only improves the accuracy of power output prediction but also streamlines the computational process, offering a promising alternative to traditional thermodynamic modeling in the field of energy production forecasting.

8. References

[1] Kesgin, U., & Heperkan, H. (2005). Simulation of thermodynamic systems using soft computing techniques. International journal of energy research, 29(7), 581-611.

[2] Khosravi, A. K. R. M. L. P. J., Koury, R. N. N., Machado, L., & Pabon, J. J. G. (2018). Prediction of wind speed and wind direction using artificial neural network, support vector regression and adaptive neuro-fuzzy inference system. *Sustainable Energy Technologies and Assessments*, *25*, 146-160.

[3] Eriksen, V. L. (Ed.). (2017). Heat recovery steam generator technology. Woodhead Publishing.

[4] 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.

[5] D'Haen J, Van den Poel D. Temporary staffing services: a data mining perspective. In: 2012 IEEE 12th international conference on data mining workshops; 2012.

[6] Che J, Wang J, Wang G. An adaptive fuzzy combination model based on selforganizing map and support vector regression for electric load forecasting.Energy 2012;37(1):657–64.

[7] Fushiki, Tadayoshi. "Estimation of prediction error by using K-fold cross-validation." Statistics and Computing 21 (2011): 137-146.

[8] Zhang, Nian, et al. "Gaussian process regression method for classification for high- dimensional data with limited samples." 2018 Eighth International Conference on Information science and technology (ICIST). IEEE, 2018.

[9] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific model development discussions, 7(1), 1525-1534.

[10] Yildiz, B., Bilbao, J. I., & Sproul, A. B. (2017). A review and analysis of regression and machine learning models on commercial building electricity load forecasting. Renewable and Sustainable Energy Reviews, 73, 1104-1122.

[11] Ritter, A., & Muñoz-Carpena, R. (2013). Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. Journal of Hydrology, 480, 33-45.

[12] Fast M, Assadi M, Deb S. Development and multi-utility of an ANN model for an industrial gas turbine. Appl Energy 2009;86(1):9–17.

[13] Kaya H, Tüfekci P, Gürgen FS. Local and global learning methods for predicting power of a combined gas & steam turbine. In: International conference on emerging trends in computer and electronics engineering (ICETCEE'2012), Dubai, March 24–25, 2012.

[14] Elish MO. A comparative study of fault density prediction in aspect-oriented systems using MLP, RBF, KNN, RT, DENFIS and SVR models. Artif Intell Rev 2012. http://dx.doi.org/10.1007/s10462-012-9348-9.

[15] Cleary JG, Trigg LE. K⁄: an instance-based learner using an entropic distance measure. In: 12th International conference on machine learning; 1995. p. 108–

[16] 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. https://doi.org/10.1016/j.ijepes.2014.02.027

[17] Lee JJ, Kang DW, Kim TS. Development of a gas turbine performance analysis program and its application. Energy 2011;36(8):5274–85.

[18] Improve steam turbine efficiency. http://www.iffco.nic.in/applications/ brihaspat.nsf/0/fddd5567e90ccfbde52569160021d1c8/$FILE/turbine.pdf [accessed: January 2012].

[19] Challagulla VUB, Bastani FB, Yen IL, Paul RA. Empirical assessment of machine learning based software defect prediction techniques. In: 10th IEEE international workshop on object-oriented real-time dependable systems, Sedona, 2–4 February, 2005. p. 263–70. http://dx.doi.org/10.1109/WORDS.2005.32.

[20] Liu H, Gopalkrishnan V, Quynh KTN, Ng W. Regression models for estimating product life cycle cost. J Intell Manuf 2009;20:401–8. http://dx.doi.org/ 10.1007/s10845-008-0114-4.