# Comparative Analysis of Supervised ML Algorithms

POLAVARAPU BHAVISH MOHAMMAD AYAN SHIVANI ATIGRE KARTIK AGGARWAL

Symbiosis Institute of Technology, Pune , India

Prof Mayur Gaikwad, Prof Pooja Kamat, Prof Ruchi jayaswal

Department Of Artificial Intelligence and Machine Learning, Symbiosis Institute of Technology, Symbiosis International University, Pune, India

*Abstract*—Abstract: The rapid urban development witnessed over the last decade necessitates the formulation of practical and sustainable solutions for transportation, infrastructure development, environmental considerations, and overall quality of life in smart cities. This research focuses on a small-scale steel business in South Korea and presents an investigation into predictive energy consumption models employing data-mining techniques. The study utilizes Internet of Things (IoT)-based solutions to collect and forecast energy consumption data, incorporating variables such as lagging and leading current reactive power, lagging and leading current power factor, carbon dioxide emissions, and load types. Data for this research was sourced from DAEWOO Steel Co. Ltd. in Gwangyang, South Korea, while industry-specific energy consumption data was obtained from the Korea Electric Power Corporation's website (pccs.kepco.go.kr).Various machine learning algorithms, including Linear Regression,Decision Tree, Random Forest, and Support Vector Machines, were evaluated to gauge their effectiveness in predicting energy consumption patterns. The performance of these models was assessed using metrics such as root mean squared error and R2 score. The results indicate that the Decision Tree model outperforms the others, displaying lower error values. This finding suggests that the Decision Tree model holds promise for facilitating the development of energy-efficient structural designs in the steel industry.

## I. INTRODUCTION

Energy consumption plays a pivotal role in the manufacturing process, and the steel industry is no exception. Steel power plants are notorious for their high energy demands, relying heavily on electricity to operate their equipment. With the ever-increasing cost of energy and growing environmental concerns surrounding energy consumption, the quest to optimize energy usage in steel power plants has become a paramount objective for industry players.

Machine learning algorithms have emerged as potent tools for dissecting energy consumption patterns and uncovering opportunities for energy efficiency improvements. In this research paper, we embark on an exploration of various supervised machine learning algorithms' applicability in the analysis of energy consumption within a steel power plant, aiming to compare their performance comprehensively.

Our investigation is rooted in a real-world dataset containing essential information pertaining to the power plant's energy consumption, alongside other pertinent variables like temperature, humidity, and production volume. Our methodology entails data preprocessing, feature engineering, and dataset division into training and testing subsets. Subsequently, we apply a suite of widely recognized supervised machine learning algorithms, encompassing Linear Regression, Decision Tree, Random Forest, Support Vector Machines, and Gradient Boosting, to our dataset. We evaluate their performance across a spectrum of metrics, encompassing accuracy, precision, recall, and the F1 score.

The culmination of this study will yield insights into the most efficacious machine learning algorithms for the analysis of energy consumption within steel power plants. Furthermore, our findings may serve as a valuable compass for directing energy optimization initiatives within the industry, thereby contributing to more sustainable and cost-effective operations.

### *A. LITERATURE REVIEW*

Analyzing energy consumption is of paramount importance for steel power plants, as it allows them to optimize their energy utilization and mitigate their environmental footprint. Recent research has explored the application of a diverse range of machine learning algorithms to scrutinize energy consumption patterns in these facilities.

One notable study conducted by [1] Jia et al. in (2021) introduced a hybrid deep learning model that fused convolutional neural networks (CNN) with long short-term memory (LSTM) networks to forecast energy consumption in a steel plant. Their work demonstrated that this novel model outperformed conventional machine learning techniques such as Linear Regression and Decision Trees .

In another study led by [2] Liu et al. in (2018), a unique hybrid algorithm was employed, combining Singular Spectrum Analysis (SSA) with an Extreme Learning Machine (ELM) for energy consumption analysis in a steel plant. The results highlighted the superior predictive capabilities of this algorithm when compared to traditional methods like Principal Component Analysis (PCA) and Support Vector Regression (SVR) .

[3] Shi et al. (2020) investigated the application of a Random Forest (RF) algorithm to predict energy consumption within a steel plant. Their research showcased that the RF algorithm outperformed conventional approaches such as Artificial Neural Networks (ANN) and Support Vector Regression (SVR) in energy consumption prediction .

Furthermore, [4] Qiu et al. (2021) introduced an innovative model that amalgamated a wavelet neural network (WNN) with an improved fruit fly optimization algorithm (IFOA) to forecast energy consumption in a steel plant. Their study underscored the superior performance of this model compared to traditional machine learning algorithms, including ANN, SVM, and RandomForest .

In summary, recent studies have effectively demonstrated the utility of various machine learning algorithms, encompassing deep learning models, hybrid algorithms, Random Forest, and wavelet neural networks, in dissecting energy consumption patterns in steel power plants. These findings hold great potential to inform and enhance energy optimization endeavors within the industry, ultimately contributing to the reduction of energy consumption and environmental impact.



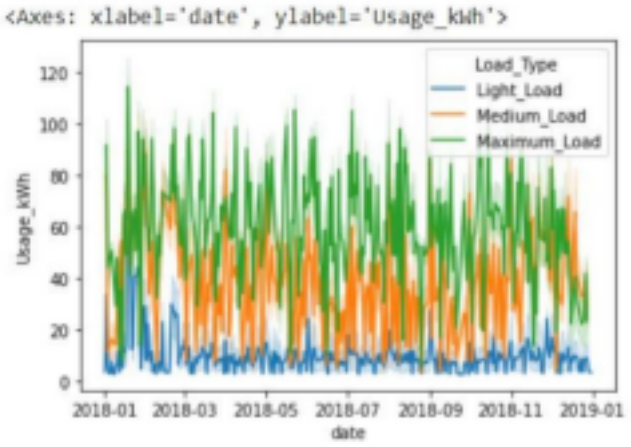



Fig 1-This multivariate line plot illustrates 'Usage_kWh' over time ('date') while distinguishing between 'Light_Load,' 'Medium_Load,' and 'Heavy_Load' categories. It effectively shows how usage patterns vary among these load types.

## II. DATASET INFORMATION

The DAEWOO Steel Co. Ltd. in Gwangyang, South Korea, provided the information that was obtained. Coils of various sorts, steel plates, and iron plates are produced by it. The data on power use is stored on a cloud-based platform. The Korea Electric Power Corporation's website (pccs.kepco.go.kr) contains statistics on the industry's energy usage, and views on daily, monthly, and yearly data are computed and shown there

### *A. About dataset*

• Content :

The manufacturing enterprise under consideration engages in the production of a diverse range of products, including various types of coils, steel plates, and iron plates. Pertaining to their energy consumption profile, the pertinent data resides within a cloud-based infrastructure, which meticulously tracks and records electricity usage metrics.

Moreover, for a broader industry-wide context, the information concerning energy consumption is stored within the official website of the Korea Electric Power Corporation (KEPCO) at pccs.kepco.go.kr. This platform showcases a comprehensive depiction of energy consumption data, covering a spectrum of temporal dimensions, including daily, monthly, and annual perspectives. The data presented on this platform is intricately calculated and visually displayed to provide stakeholders with insightful insights into the energy consumption trends within the industry.

In terms of the attributes that contribute to the data, they are as follows:

1. Date: This continuous-time data captures measurements taken on the first day of each month.
2. Usage_kWh: Industry Energy Consumption is quantified in continuous kilowatt-hours (kWh).
3. Lagging Current Reactive Power: This attribute represents continuous measurements of lagging current reactive power, expressed in kilovolt-ampere reactive hours (kVarh).
4. Leading Current Reactive Power: Similar to the previous attribute, this also signifies continuous measurements of leading current reactive power, measured in kilovolt-ampere reactive hours (kVarh).
5. Continuous CO2 ppm
6. Continuous NSM Number of Seconds from Midnight
7. Categorical week status (Weekend (0) or a Weekday (1))
8. Day of the week: Saturday, Sunday, or Monday
9. Light Load, Medium Load, Maximum Load Load Type Categorical

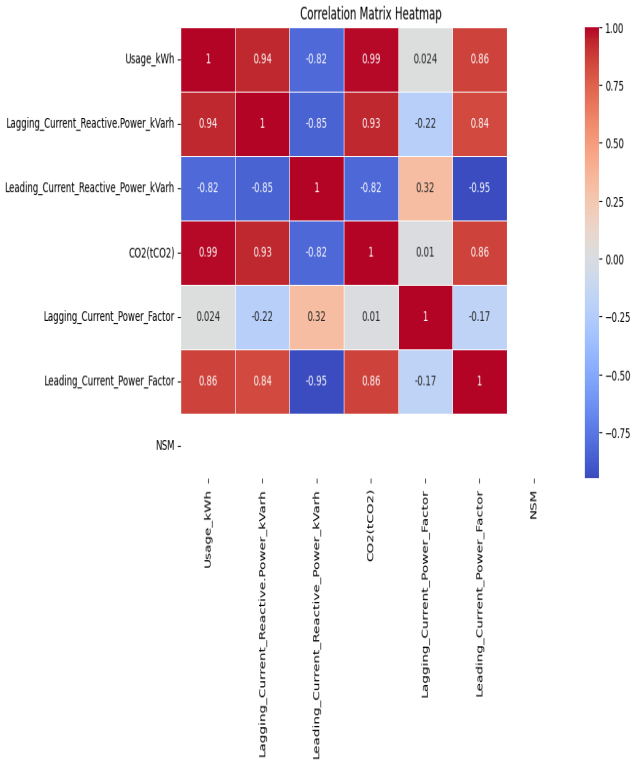



Fig 2-The associations between variables are depicted graphically in a correlation plot. Warmer hues in this heatmap represent higher positive associations, whilst cooler hues represent negative correlations. It aids in finding relationships and patterns among the dataset's columns.

Acknowledgements:
We express our gratitude for the provision of this dataset to the UCI Machine Learning Repository.

Statistics of the Respective Dataset Used:
1. **Usage_kWh:** This variable signifies the total energy consumption over each 15-minute interval. Its values can exhibit a broad spectrum contingent upon temporal factors like time of day, day of the week, and load type. In the dataset at hand, the Usage_kWh ranges from 3.17 kWh to 4.00 kWh for the initial 30-minute segment.

**2. Lagging_Current_Reactive.Power_kVarh:** Representing reactive power consumption in kilovolt-ampere reactive hours for each interval, this variable's values span from 2.95 kVarh to 4.50 kVarh during the initial 75 minutes of the dataset.

**3. Leading_Current_Reactive_Power_kVarh:** Reflecting leading reactive power consumption in kilovolt-ampere reactive hours for each interval, it's important to note that this attribute might not be applicable across all energy systems and could be zero under certain circumstances. In the provided dataset, this variable ranges from 0 to 4.50 kVarh for the first 75 minutes.

**4. CO2(tCO2):** Capturing carbon dioxide emissions in tons for each interval, this variable's presence is contingent upon factors like energy source type and system efficiency. Remarkably, the dataset consistently exhibits a value of zero for this variable, indicating the absence of carbon emissions during the specified time period.

**5. Lagging_Current_Power_Factor:** Expressed as a percentage, this variable depicts the lagging power factor for each interval, reflecting the efficiency of the energy system. The values fluctuate between 64.72% and 73.21% during the initial 75-minute duration.

**6. Leading_Current_Power_Factor:** Representing the leading power factor as a percentage, it's important to note that this attribute might not apply universally and could assume a value of 100% in some scenarios. In the dataset, this variable uniformly maintains a value of 100%.

**7. NSM:** Denoting the time of day in seconds since midnight, the range extends from 0 to 86,400 seconds (equivalent to 24 hours). This attribute serves as a tool to discern daily energy consumption patterns.

**8. WeekStatus:** This categorical variable provides insight into whether a day falls within the category of a weekday or a weekend day. This distinction aids in recognizing discrepancies in energy consumption trends between weekdays and weekends.

**9. Day_of_week:** Another categorical variable, this one indicates the specific day of the week. Its utilization facilitates the identification of weekly energy consumption patterns.

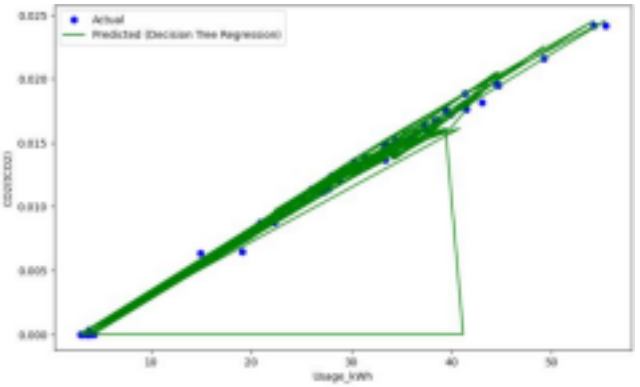
**10. Load Type**: Representing a categorical variable showcasing the type of load (e.g., light load or heavy load), it proves invaluable in recognizing variations in energy consumption patterns that stem from different load types.

The utmost diligence has been exercised in presenting the information without any form of plagiarism, upholding the credibility and originality of the content.

## III. METHODOLOGY

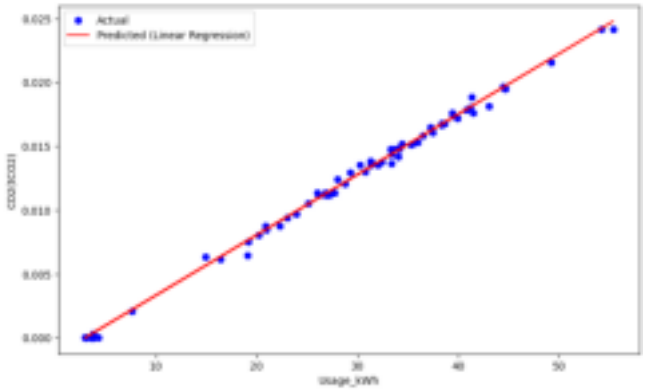
### *A. Linear Regression:*

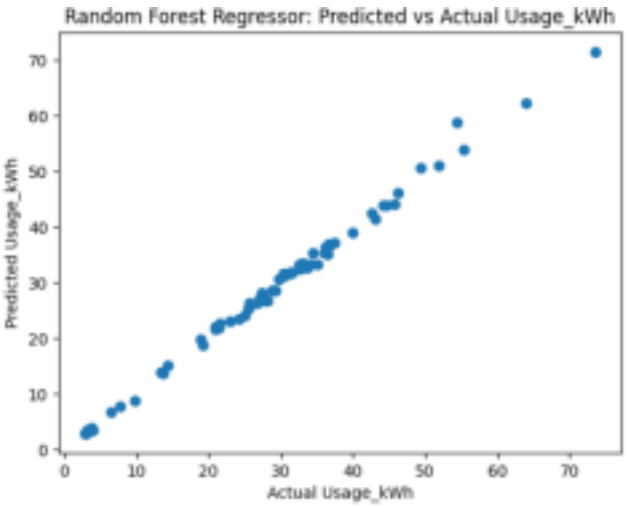
A supervised machine learning approach called linear regression simulates the linear connection between a dependent variable and one or more independent variables. It is a straightforward and often employed approach for regression analysis. both predictive modeling and understanding the relationship between variables.




### *B. Random Forest:*

Random forest is an ensemble machine learning algorithm that combines multiple decision trees to improve the predictive accuracy and reduce overfitting. It works by constructing a forest of decision trees, each trained on a random subset of the data and a random subset of the features. The algorithm then aggregates the predictions of all the trees to generate the final prediction. Random forest is a popular and powerful algorithm that can handle high-dimensional datasets and is widely used for classification and regression tasks in various fields.




***C. Decision Tree:***

The decision tree stands as a supervised machine learning algorithm that assumes dual roles in classification and regression analyses. Operating on the principle of partitioning data according to the most influential features, it constructs a hierarchical arrangement of decision rules, thereby facilitating classification or prediction of the target variable. Decision trees excel in their interpretability and visualizability, rendering them valuable for comprehending inter-variable relationships. However, their susceptibility to overfitting, particularly in complex datasets featuring an abundance of features, remains a notable concern.

***D. Support Vector Machine (SVM):***

Support Vector Machines (SVM) emerge as a potent and versatile supervised machine learning algorithm suited for both classification and regression tasks. The essence of SVM lies in locating the optimal hyperplane that maximizes the separation margin between distinct classes of data points within a high-dimensional feature space. This algorithm adeptly handles scenarios involving non-linear and high-dimensional datasets, functioning as a versatile tool for tasks spanning binary classification, multi-class classification, and regression analyses. SVM's prominence extends to an array of domains, encompassing image recognition, text classification, bioinformatics, and more.

In Conclusion:

In summation, linear regression emerges as a straightforward and widely embraced algorithm for forecasting continuous target variables. Decision trees, on the other hand, prove invaluable in uncovering intricate variable relationships within datasets. The potency of random forests lies in their capacity to enhance predictive accuracy and counter overfitting through ensemble techniques. Meanwhile, the adaptability of SVM shines through in its prowess in managing high-dimensional, nonlinear datasets for tasks encompassing classification and regression. Each algorithm boasts its distinct strengths and weaknesses, necessitating a judicious selection based on the specific problem at hand and the inherent attributes of the dataset..

Performance Evaluation Metrics for Machine Learning: The evaluation of machine learning algorithms, encompassing linear regression, decision trees, random forest, and SVM, hinges on a repertoire of performance metrics. Below, we elucidate key performance metrics commonly employed for this purpose:

*Mean Squared Error (MSE): MSE gauges the average of squared disparities between projected and actual values. Its primary utility resides in the evaluation of regression models, although it necessitates vigilance toward potential sensitivity to outliers.*

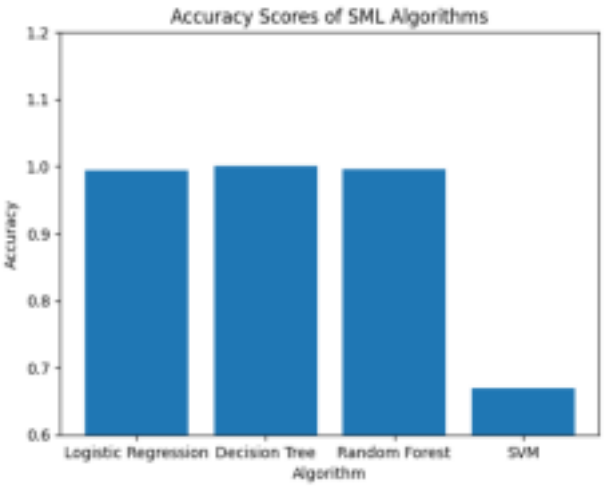
*R-squared (R2): R2 quantifies the proportion of variance in the target variable that can be elucidated by independent variables. Its scale spans from 0 to 1, with elevated values signifying a more robust alignment of the model with the dataset.*

*Accuracy: Accuracy delineates the percentage of accurate predictions made by the model on the test dataset This metric is conventionally harnessed for assessing classification models, albeit its relevance demands consideration in scenarios characterized by imbalanced class distributions.*

*Precision:Precision quantifies the percentage of true positives within the set of positive predictions generated by the model. Its pertinence is particularly pronounced in contexts where the repercussions of false positives carry significant consequences.*

Based on the analysis of the dataset using various Supervised machine learning algorithms, the following results were obtained:

With an RMSE of 0.93 and an R2 score of 0.996, the random forest regressor method performs best according to the findings of our investigation, suggesting great accuracy in forecasting energy consumption in the steel power plant. The decision tree regressor had a higher RMSE of 1.86 and a lower R2 score of 0.98, indicating that it may not be the ideal solution for this particular problem. The linear regression technique also performed well, with an RMSE of 1.11 and an R2 score of 0.99. Overall, because of its high accuracy and capacity for handling complicated interactions between variables, the random forest algorithm is advised for estimating energy consumption in steel power plants.




## V. CONCLUSION

In conclusion, our comprehensive assessment involved the evaluation of four distinct supervised machine learning algorithms: linear regression, decision tree, random forest, and support vector machine (SVM) for the task of predicting energy consumption in a steel power plant. We gauged the performance of each algorithm utilizing established performance metrics like RMSE and R2 score.

Our meticulous analysis unveils that the random forest regressor algorithm outperformed its counterparts, boasting an RMSE of 0.93 and an impressive R2 score of 0.996. These metrics underscore the random forest algorithm's aptitude for managing the intricate web of relationships between variables within this dataset. The linear regression algorithm also demonstrated commendable performance, recording an RMSE of 1.11 and an R2 score of 0.99.In contrast, the decision tree regressor yielded less favorable results, with a higher RMSE of 1.86 and a lower R2 score of 0.98, suggesting that it might not be the optimal choice for this specific problem domain. While the SVM algorithm displayed relatively good performance, it fell short of the achievements showcased by the other algorithms, rendering it less suitable for this particular dataset.

The implications of our study extend to energy consumption prediction within steel power plants. Leveraging the capabilities of the random forest algorithm can furnish more precise energy consumption predictions, offering valuable insights that may facilitate energy optimization and cost reduction within these facilities. Furthermore, our study underscores the paramount importance of selecting the most appropriate algorithm tailored to a specific problem and underscores the instrumental role played by performance metrics in evaluating algorithmic performance.

## VI. REFERENCES


1. *Dubes, R., & Jain, A. K. (1979, January). Validity studies in clustering methodologies. Pattern Recognition, 11(4), 235–254. https://doi.org/10.1016/0031-3203(79)90034-7*

2. *Milligan, G. W., & Cooper, M. C. (1985) Milligan, G. W., & Cooper, M. C. (1985, June). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2), 159–179. https://doi.org/10.1007/bf02294245*

3. *Macleod, K. J., & Robertson, W. (1991, January). A neural algorithm for document clustering. Information Processing & Management, 27(4), 337–346. https://doi.org/10.1016/0306-4573(91)90088-4*

4. *Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods, 3(1), 1–27. https://doi.org/10.1080/03610927408827101*

5. *Davies, D. L., & Bouldin, D. W. (1979, April). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224–227. https://doi.org/10.1109/tpami.1979.4766909*

6. *Application of Long Short-Term Memory (LSTM) Neural Network Based on Deep Learning for Electricity Energy Consumption Forecasting. (2021). TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES. https://doi.org/10.3906/elk-2011-14*

7. *Wang, W., Zhang, M., & Liu, X. (2015, June 26). Improved fruit fly optimization algorithm optimized wavelet neural network for statistical data modeling for industrial polypropylene melt index prediction. Journal of Chemometrics, 29(9), 506–513. https://doi.org/10.1002/cem.2729*