EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Electricity Consumption Classification using Various Machine Learning Models

Bijay Kumar Paikaray[1,*] Swarna Prabha Jena[2], Jayanta Mondal[3], Nguyen Van Thuan[4], Nguyen Trong Tung[5], Chandrakant Mallick[6]

[1]Centre for Data Science, Department of Computer Science & Engineering, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India
[2]Department of Electronics and Communication Engineering, Centurion of University of Technology and Management, Odisha, India
[3]School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India
[4]Faculty of Engineering Technology, Hung Vuong University, Vietnam
[5]Quality Assurance, Dong A University, Vietnam
[6]Department of Computer Science & Information Technology, GITA Autonomous College, Bhubaneswar, Odisha, India

Abstract

INTRODUCTION: As population has increased over successive generations, human dependency on electricity has increased to the point where it has become a norm and indispensable, and the idea of living without it has become unthinkable.
OBJECTIVES: Machine learning is emerging as a fundamental method for performing tasks autonomously without human intervention. Forecasting electricity consumption is challenging due to the many factors that influence it; embracing modern technology with its heavy focus on machine learning and artificial intelligence is a potential solution.
METHODS: This study employs various machine learning algorithms to forecast power usage and determine which method performs best in predicting the dataset based on different variables.
RESULTS: Eight models were tested, including Linear Regression, DT Classifier, RF Classifier, KNN, DT Regression, SVM, Logistic Regression, and GNB Classifier. The Decision Tree model had the greatest accuracy of 98.3%.
CONCLUSION: The Decision Tree model's accuracy can facilitate efficient use of electricity, leading to both conservation of electricity and cost savings, and be a guiding light in future planning.

Keywords: Electricity Prediction, Machine Learning, SkLearn.

Received on 25 02 2024, accepted on 18 05 2024, published on 07 06 2024

doi: 10.4108/ew.6274

1. Introduction

Since the time power of electricity first began to become harnessed for civilian use, it has increasingly become an important human asset. As population has increased over successive generations, human dependency on electricity has increased until it has become a norm, and the thought of living without electricity has become unthinkable. Such is the case in today's world, where such a dependency brings crisis when electricity is cut off even for a short time, which can happen for a variety of uncertain reasons.

The gradual development of human society has led to an increasing reliance on electricity. Initially, electricity was mostly utilized for work-related tasks. It was then regarded as a privilege exclusive to affluent individuals. Over time, energy consumption became diversified in definition and application, and continued to rise steadily.

Efficient electricity usage requires systematic development and robust energy management. It is crucial to develop a detailed plan for managing electricity use, taking into account elements such as how, where, and when electricity is used. This is where predictive models for power consumption,

[*]Corresponding author. Email: bijaypaikaray87@gmail.com

utilizing machine learning techniques, can address the issue of wasteful electricity usage.

By reviewing the current literature, we examined current answers to this difficult research challenge. Section 3 provides a conceptual overview of the suggested approach, Section 4 provides a concise overview of the simulation, including screenshots of test outputs from the application. The study's future prospects are outlined in section 5.

2. Literature Review

In developing this electricity prediction model approximately 20 research papers were reviewed. Numerous studies have been conducted, each differing primarily in the datasets and algorithms used. The factors used as features and the algorithms themselves are distinct across these studies. Algorithms utilized include XGboost, long short-term memory networks, artificial neural networks, and support vector machines, amongst others. The accuracy of these models varies due to differences in the dataset and features used.

Table 1: Algorithms used in the referenced paper.

Ref. No.	Algorithms	Ref. No.	Algorithms
1	PSO (iPSO), PCA, iPSO-ANN, GA-ANN, CPSO, GA-ANFIS	11	XGBoost
2	NNGM, AGM, BPN, SVR, MAPE, GPRM, SLP, GM	12	SVR
3	ANN, HVAC,	13	ANN
4	SVR, MAED, TPES, MENR, SPO, GPRM, GNP, SVC,	14	Bi-LSTM
5	ARIMA, ROLLING-ALO-GM, NOGM	15	ANN
6	SVM, CCC, RPE, GHG	16	SVR
7	XGboost, FDN	17	SVR
8	LSTM	18	LSTM
9	TLBO-BP, TLBO-SVM, PSO, GA, SVM, ANN, ELM, AdaBoost, ELM, TLBO	19	E-LSTM
10	GM (1,1), IRGM, PSO, TRF, OICGM, ARIMA	20	XGBoost, SARIMA

3. Methodology

The first step in this work involved collecting a dataset from GitHub. The discrete dataset contains 29 features and 2,980 rows of data. Data wrangling, an essential process which greatly influences the outcome of any prediction model, was meticulously performed. The process must be done very carefully, as improper data wrangling can lead to data type errors and the model's inability to process the data correctly. Following data wrangling, the data preparation process was carried out to set up the dataset, as well as the platform for model building.

3.1. Hardware Used

The hardware configurations include compatibility with the AMD Ryzen 5 and Intel Core i5 processors. The device features a Radeon Vega Mobile Gfx GPU running at 2.10 GHz or an NVIDIA GeForce RTX 3050 with 8GB of RAM, 512GB of ROM, and a display refresh rate of 165Hz/60Hz. This hardware enabled us to host the implementation of the Machine Learning model.

3.2. Software Used

Anaconda: An open-source distribution of the Python and R programming languages, Anaconda is used for computing and solving numerous machine-learning algorithms. The platform aids in overseeing packaging and deployment and allows for the installation and utilization of several programs, including NumPy, SciPy, and Matplotlib. The platform supports the validation and advancement of various machine learning and artificial intelligence applications.

Jupyter Notebook: This tool enables users to solve equations, create visualizations, and incorporate narrative prose. It is a fundamental tool extensively utilized for data analysis, scientific research, and machine learning. The electricity consumption forecast model was developed using Jupyter Notebook after completing the data preparation phase. Multiple machine learning methods were utilized to develop the prediction model, and their effectiveness was evaluated based on the dataset.

3.3. Algorithms

For this work seven algorithms were used: Linear Regression (LR), Decision Tree (DT) Classifier, Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree Regression (DT Regression), Support Vector Machine (SVM), and Gaussian Naïve Bayes (GNB) Classifier.

Linear Regression (LR): Linear Regression is one of the simplest and most popular machine learning models for prediction. It makes predictions for continuous or numeric data, showing a linear relationship between two variables: the dependent variable (Y) and one or more independent variables (X). X is directly proportional to Y; the relationship is expressed by the equation (1).

$$Y = mX + c \qquad (1)$$

where m and c are parameters initially assigning random values. The cost function, typically the mean squared error, is used to optimize these parameters.

Decision Tree (DT) Classifier: Appropriate for both classification and regression problems, the Decision Tree algorithm structures data in a tree-like model of decisions, where core nodes represent features, branches represent decision rules and the leaves represent outcomes. This structure visually represents all potential solutions to a problem under specified conditions, as depicted in equation (2). The process begins with the dataset, determining a tree structure and decision rules at each node.

$$D = \{X, y\} \qquad (2)$$

Random Forest (RF) Regression: This supervised machine learning strategy utilizes several decision trees trained on various subsets of the dataset. Aggregating the average of all decision trees enhances the accuracy of predictions. Having 'more trees in the forest' leads to an increased accuracy and helps prevent overfitting.

K- Nearest Neighbour: KNN uses proximity to make predictions about the grouping of an individual data point. The distance between two points is calculated using the Euclidean distance formula (3):

$$d = \sqrt{((x_2 - x_1)^2 - (y_2 - y_1)^2)} \qquad (3)$$

Support Vector Machine: This supervised machine learning method is appropriate for regression and classification tasks. It operates by converting data into a high-dimensional feature space to categorize data points, even if a straight line cannot separate them. Equation (4) denotes the optimal hyperplane for the data points.

$$W.x + b = 0 \qquad (4)$$

Logistic regression (LR): Logistic Regression is a statistical method for building ML models where the dependent variable is dichotomous. There are three logistic regressions: Binary, Multinomial, and Ordinal. The equation is (5):

$$Y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \qquad (5)$$

In Logistic Regression ranges, Y from 0 to 1. To achieve this, the equation can be divided by (1 - Y) resulting in Y/(1-Y), where Y = 0 yields 0 and Y = 1 yields infinity. To obtain a range from negative infinity to positive infinity, we take the logarithm of the equation, resulting in (6).

$$\log(y/(1-y) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \qquad (6)$$

Gaussian Naive Bayes (GNB): Gaussian Naive Bayes is a technique used in ML based on the probabilistic approach and Gaussian distribution, which produces a bell-shaped curve. It makes predictions based on the training data provided to the model. This method is very simple and is used for real-time applications. The principles can be expressed using the following notations represented in equations (7) and (8):

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad (7)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad (8)$$

3.4 Libraries

Various libraries were required for the predictive model: the following packages were installed. Table 2 highlights the contributions of each library or package.

Table 2. Python Libraries and their uses.

Name	Application
Numpy	Used for mathematical operations on an array.
Pandas	Used for data analysis.
Matplotlib	Used for creating visualizations for the users understanding.
Seaborn	Used to provide graphics.
Pingouin	Used for statistics Operations.
Sklearn	Helped implement machine learning models and simulate statistical modelling.

4. Result and Discussion

Linear Regression Model

The variables on the X and Y coordinates closely align with the straight line (Fig 1), indicating better data and analysis within this model, with a calculated accuracy of 92.17%.

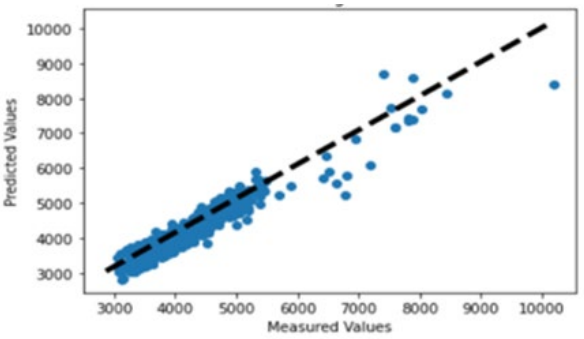


Fig. 1. Linear Regression Model

Random Forest Regression Model

The results for the random forest regression model when applied to the dataset is shown in Fig 2. It reveals that the given parameters work effectively with the model, achieving an accuracy of 95.95%.

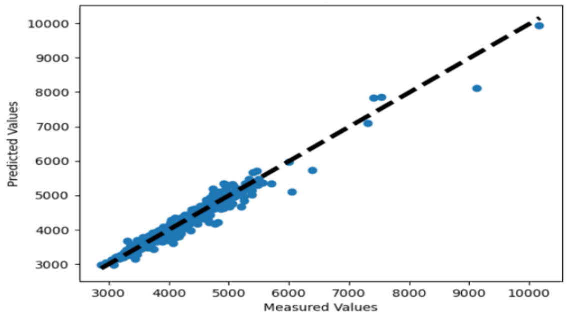


Fig. 2. Random Forest Regression Model

Decision Tree Regression Model

Results indicate that the given parameters also work effectively with this decision tree regression, yielding an accuracy of 91.27% (Fig 3).

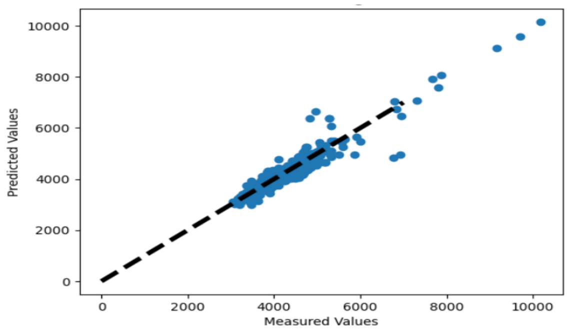


Fig. 3. Decision Tree Regression Model

K-Nearest Neighbour

Using this algorithm, we structured a confusion matrix. The predicted classes are on the X axis and actual classes on the Y axis, illustrating the relationship. The matrix shows 116 classes meet the conditions, while 11 do not. In the 1-0 structure, 7 classes satisfy the conditions, and in 1-1 format, 611 do not (Fig 4). This algorithm achieved an accuracy score of 97.58%, indicating a high level of accuracy.

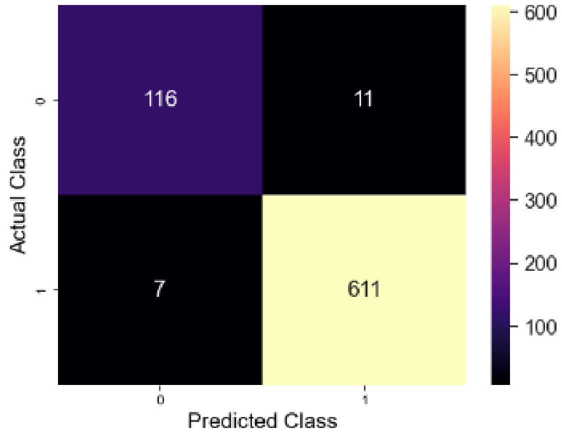
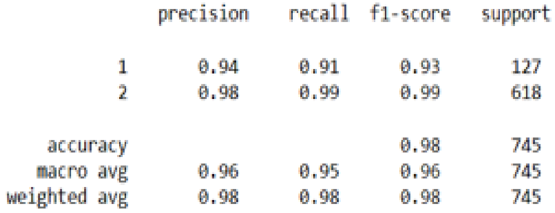


Fig. 4. Confusion Matrix for KNN algorithm to our project work.



```
              precision    recall  f1-score   support

           1       0.94      0.91      0.93       127
           2       0.98      0.99      0.99       618

    accuracy                           0.98       745
   macro avg       0.96      0.95      0.96       745
weighted avg       0.98      0.98      0.98       745
```

Fig. 5. Classification report of KNN algorithm to our project work.

Logistic Regression

Using this algorithm, we constructed a confusion matrix with the predicted class on the X axis, and actual class on the Y axis. The matrix illustrates the relationship between these two classes. Out of 83 categories, most results are satisfactory, but some are not. This algorithm yielded among all the algorithms we tested, achieving only 86.30%, which is not a well-developed model for our project needs. A classification report was generated to validate the data, as shown in Fig 7.

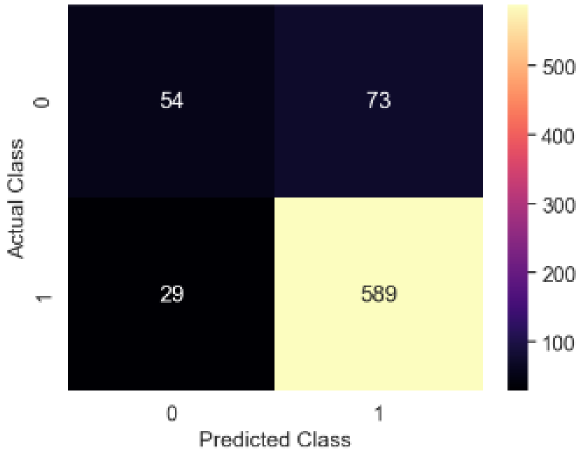
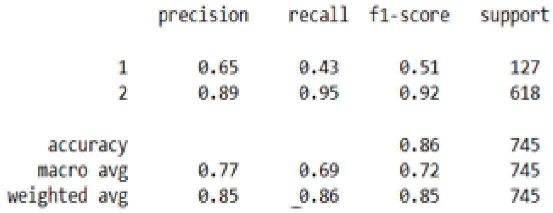


Fig. 6. Confusion Matrix for Logistic algorithm



```
              precision    recall  f1-score   support

           1       0.65      0.43      0.51       127
           2       0.89      0.95      0.92       618

    accuracy                           0.86       745
   macro avg       0.77      0.69      0.72       745
weighted avg       0.85     _0.86      0.85       745
```

Fig. 7. Classification report of Logistic algorithm

Decision tree Classification Model

Using this algorithm, a confusion matrix was structured with the predicted class on the X-axis, and the actual class on the Y-axis, illustrating the relationship between the two classes. In the first class, 121 classes are satisfied, with six unsatisfied. In the second class, 612 were satisfied, and six unsatisfied, as shown in Fig 8. This algorithm attained the highest accuracy, achieving a score of 98.38 %. Among all the algorithms used in our project, this one proved to be the most accurate and reliable.

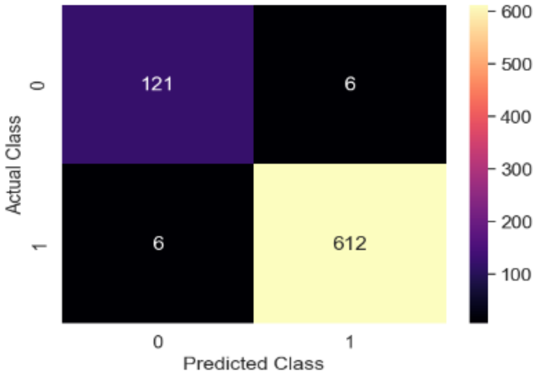
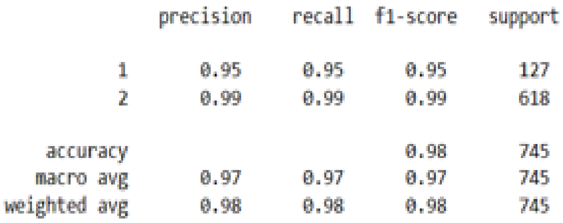


Fig. 8. Confusion Matrix for Decision tree classification report



```
              precision    recall  f1-score   support

           1       0.95      0.95      0.95       127
           2       0.99      0.99      0.99       618

    accuracy                           0.98       745
   macro avg       0.97      0.97      0.97       745
weighted avg       0.98      0.98      0.98       745
```

Fig. 9. Report of the Decision tree classification report

Support Vector Machine

Using this algorithm, we produced a confusion matrix with the predicted class on the X-axis and the actual class on the Y-axis, showing the relationship between these two classes. In this first class, 117 are satisfied with this model, and 10 unsatisfied. In the second class, 615 are satisfied, with three un satisfied (Fig 10).

This algorithm achieved the second highest accuracy for our model, with a 98.25 % score (Fig 11). Among all the models used, this remains one of the most accurate.

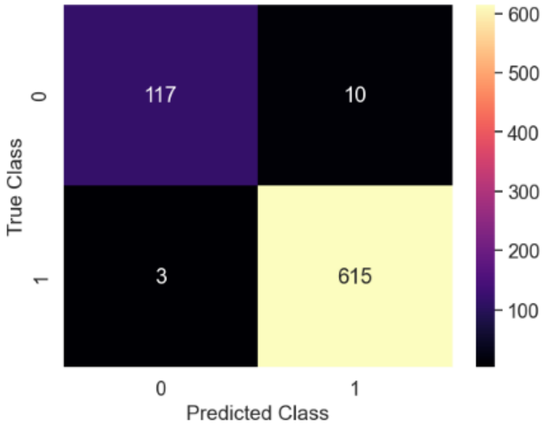
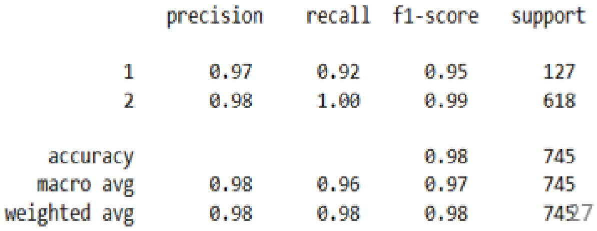


Fig. 10. Confusion Matrix for SVM Classification.



```
              precision    recall  f1-score   support

           1       0.97      0.92      0.95       127
           2       0.98      1.00      0.99       618

    accuracy                           0.98       745
   macro avg       0.98      0.96      0.97       745
weighted avg       0.98      0.98      0.98       74527
```

Fig. 11. Report of the SVM Prediction model report

Gaussian NB Model

Using the Gaussian Naive Bayes model, we constructed a confusion matrix with the predicted class on the X-axis and actual class on the Y-axis. The matrix demonstrates the relationship between the two categories. Out of 127 classes in the first session, 123 were correctly classified, while four were not. In the second class, 570 individuals were correctly classified, while 48 were misclassified, according to Figure 12. The algorithm achieved an accuracy score of 97.58 %, indicating high performance, As shown in Figure 13.

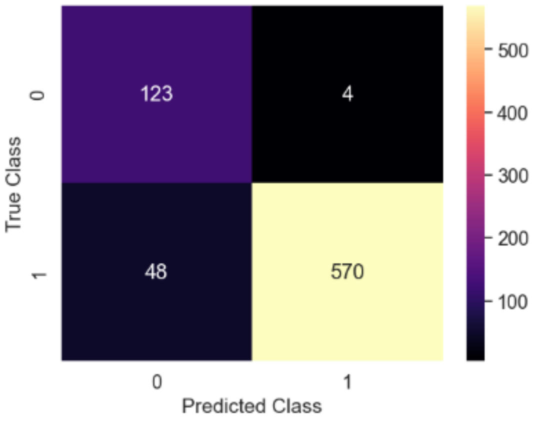
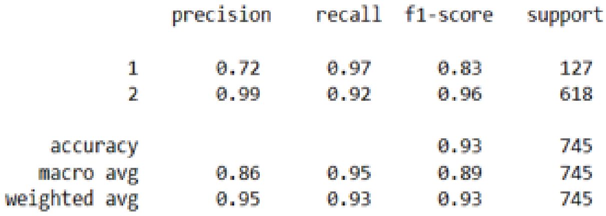


Fig. 12. Confusion Matrix for Gaussian NB Classification



```
              precision    recall  f1-score   support

           1       0.72      0.97      0.83       127
           2       0.99      0.92      0.96       618

    accuracy                           0.93       745
   macro avg       0.86      0.95      0.89       745
weighted avg       0.95      0.93      0.93       745
```

Fig. 13. Report of the Gaussian NB Prediction Model

Table 3 shows the accuracy obtained for each machine learning algorithm on the dataset.

Table 3. Algorithms used in the referenced paper.

Algorithms	Accuracy (%)
Decision Tree Classification	98.3
Support Vector Machine	98.2
KNN	97.5
Gaussian NB	97.5
Random Forest Regression	95.9
Linear Regression	92.1
Decision Tree Regression	91.2
Logistic Regression	86.3

5. Conclusion and Future Scope

The study concluded that there are a variety of machine learning algorithms which can be applied to the electricity consumption dataset with significant results to increase efficiency, including LR, DT Classifier, KNN, Decision Tree Regression, SVM, Logistic Regression, and Gaussian Naive Bayes Classifier. Among these models, the Decision Tree algorithm achieved the highest accuracy achieved was 98.3. This model can be utilized in various applications for both casual and industrial purposes. The application of the model can assist in laying out plans for the efficient use of electricity, leading to conservation of both electricity and money.

References

1. Li, Kangji, *et al.* "Building's electricity consumption prediction using optimized artificial neural networks and principal component analysis." Energy and Buildings 108 (2015): 106-113.
2. Hu, Yi-Chung. "Electricity consumption prediction using a neural-network-based grey forecasting approach." Journal of the Operational Research Society 68 (2017): 1259-1264.
3. Beccali, M., *et al.* "Short-term prediction of household electricity consumption: Assessing weather sensitivity in a Mediterranean area." Renewable and Sustainable Energy Reviews 12.8 (2008): 2040-2065.
4. Kavaklioglu, Kadir. "Modeling and prediction of Turkey's electricity consumption using Support Vector Regression." Applied Energy 88.1 (2011): 368-375.
5. Ding, Song, Keith W. Hipel, and Yao-guo Dang. "Forecasting China's electricity consumption using a new grey prediction model." Energy 149 (2018): 314-328.
6. Shine, P., *et al.* "Annual electricity consumption prediction and future expansion analysis on dairy farms using a support vector machine." Applied energy 250 (2019): 1110-1119.
7. Chen, Kunlong, *et al.* "A novel data-driven approach for residential electricity consumption prediction based on ensemble learning." Energy 150 (2018): 49-60.
8. Lin, Zhifeng, Lianglun Cheng, and Guoheng Huang. "Electricity consumption prediction based on LSTM with attention mechanism." IEEJ Transactions on Electrical and Electronic Engineering 15.4 (2020): 556-562.
9. Li, Kangji, *et al.* "Short-term electricity consumption prediction for buildings using data-driven swarm intelligence-based ensemble model." Energy and Buildings 231 (2021): 110558.
10. Xu, Ning, Yaoguo Dang, and Yande Gong. "Novel grey prediction model with nonlinear optimized time response method for forecasting of electricity consumption in China." Energy 118 (2017): 473-480.
11. W Wang, Y Shi, G Lyu, & W Deng (2017). Electricity consumption prediction using XGBoost based on discrete wavelet transform. DEStech Transactions on Computer Science and Engineering.
12. Kadir Kavaklioglu (2011). Modeling and prediction of Turkey's electricity consumption using Support Vector Regression. Applied Energy, 88(1), 368-375.
13. Lambros Ekonomou (2010). Greek long-term energy consumption prediction using artificial neural networks. Energy, 35(2), 512-517.
14. Ghosh, H., Rahat, I. S., Mohanty, S. N., & Ramesh, J. V. N. (2023). Microbial Image Deciphering: Navigating Challenges with Machine and Deep Learning.
15. Platon, R., Dehkordi, V. R., & Martel, J. (2015). Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis. Energy and Buildings, 92, 10-18.
16. Magoulès, F., Piliougine, M., & Elizondo, D. (2016, August). Support vector regression for electricity consumption prediction in a building in japan. In 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES) (pp. 189-196). IEEE.
17. Sun, S., & Chen, H. (2021, August). Data-driven sensitivity analysis and electricity consumption prediction for water source heat pump system using limited information. In Building Simulation (Vol. 14, pp. 1005-1016). Tsinghua University Press.
18. Jena, S. P., Paikaray, B. K., Pramanik, J., Thapa, R., & Samal, A. K. (2023). Classifications on wine informatics using PCA, LDA, and supervised machine learning techniques. International Journal of Work Innovation, 4(1), 58-73.
19. Hora, S. K., Poongodan, R., De Prado, R. P., Wozniak, M., & Divakarachari, P. B. (2021). Long short-term memory network-based metaheuristic for effective electric energy consumption prediction. Applied Sciences, 11(23), 11263.
20. Chinnaraji, R., & Ragupathy, P. (2022). Accurate electricity consumption prediction using enhanced long short-term memory. IET Communications, 16(8), 830-844.
21. Jena, S. P., Yadav, A. K., Gupta, D., & Paikaray, B. K. (2023, September). Prediction of Stock Price Using Machine Learning Techniques. In 2023 IEEE 2nd International Conference on Industrial Electronics: Developments & Applications (ICIDeA) (pp. 169-174). IEEE.
22. Zielińska-Sitkiewicz, M., Chrzanowska, M., Furmańczyk, K., & Paczutkowski, K. (2021). Analysis of electricity consumption in Poland using prediction models and neural networks. Energies, 14(20), 6619.