Vol. 7, No. 1, February 2025, pp. 11~18, p-ISSN 2685-2381, e-ISSN 2715-2626
http://dx.doi.org/10.28989/avitec.v7i1.2722

Comparison of Machine Learning Methods for Predicting Electrical Energy Consumption

Retno Wahyusari[1]*, Sunardi[2], Abdul Fadlil[3]
[1]Departement of Informatics, Universitas Ahmad Dahlan, Indonesia
[1]Departement of Informatics, Sekolah Tinggi Teknologi Ronggolawe, Indonesia
[2, 3]Departement of Electrical Engineering, Universitas Ahmad Dahlan, Indonesia

Article Info

Article history:
Submitted December 18, 2024
Accepted January 16, 2025
Published February 3, 2025

Keywords:
CatBoost;
energy consumption;
machine learning;
data normalization;
Random Forest.

ABSTRACT

This research investigates how to accurately predict electrical energy consumption to address growing global energy demands. The study employs three Machine Learning (ML) models: k-Nearest Neighbors (KNN), Random Forest (RF), and CatBoost. To enhance prediction accuracy, the researchers included a data pre-processing step using min-max normalization. The analysis utilized a dataset containing 52,416 records of power consumption from Tetouan City. The dataset was divided into training and testing sets using different ratios (90:10, 80:20, 50:50) to evaluate model performance. Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) were used to assess prediction accuracy. Min-max normalization significantly improved KNN's performance (reduced RMSE and MAPE). RF achieved similar accuracy with and without normalization. CatBoost also demonstrated stable performance regardless of normalization. Data pre-processing, specifically min-max normalization, is crucial for improving the accuracy of distance-based algorithms like KNN. Decision tree-based algorithms like RF and CatBoost are less sensitive to data normalization. These findings emphasize the importance of selecting appropriate pre-processing techniques to optimize energy consumption prediction models, which can contribute to better energy management strategies.

Corresponding Author:
Retno Wahyusari,
Departement of Informatics, Universitas Ahmad Dahlan,
Ringroad Selatan, Yogyakarta 55191, Indonesia.
Email: *2437083005@webmail.uad.ac.id

1. INTRODUCTION

Electrical energy is a primary need in daily activities and is an important factor in the development of a country. The need for electrical energy continues to increase along with population and industrial growth[1], [2]. The efficiency of electrical energy use is a special concern due to the problem of limited resources and the need for desire. Therefore, in order to minimize waste of resources and optimize energy distribution, an accurate electrical energy consumption prediction method is needed for energy planning or management [3]. The machine learning (ML) method appears as a solution in data processing in predicting electrical energy consumption.

ML methods can improve prediction accuracy and flexibility. Algorithms such as k-Nearest Neighbors (KNN), Random Forest (RF), and CatBoost are popular algorithms used in prediction studies, this is because of their ability to handle datasets with complex features and high data variability [4], [5]. The KNN algorithm has been widely used for predictions in the fields of health, finance, education, natural disasters, and others [6]–[10]. Research conducted by Fan Li and Guang Jin [11] conducted electrical energy load prediction showing that KNN-based electrical energy load forecasting can analyze and predict electrical loads in a short time with high accuracy, as well as provide data anomaly warnings and valid and accurate data analysis. This is an advantage of the KNN algorithm, one of the simple but powerful non-parametric algorithms in classification and regression [12]–[14].

Prediction using the RF method has also been carried out in various fields [15], [16]. Ergi Putra Febtiawan, et al [17] conducted research on forecasting energy production produced by photovoltaic devices using the random forest classification method. The RF algorithm is an ensemble learning that combines several decision trees to produce more stable and accurate predictions [18], [19].

Catboost is a gradient boosting-based algorithm developed by Yandex and specifically designed to handle categorical features efficiently without requiring one-hot encoding [20], [21]. Abdullahi A. Ibrahim et al. [22] compared the Catboost algorithm with other ML methods in predicting loan applications and staff promotions, producing good performance with a precision value of 0.83.

Dat Thanh Tran [23] stated that in building an ML model, one of the most important steps is the pre-processing process in the form of data normalization. Data normalization is the process of organizing data to minimize redundancy and maintain data integrity [24]. Research conducted by Andri Pranolo et al [25] showed that the min-max normalization method regularly obtained superior results compared to the z-score method. Min-max normalization specifically resulted in a decrease in MAPE and RMSE, as well as an increase in the R^2 value.

Based on the description above, predicting energy consumption accurately is crucial to optimizing global electrical energy usage. This study leverages advanced ML algorithms to enhance prediction performance. The prediction of electrical energy consumption uses a combination of data normalization process and KNN, RF, and CatBoost algorithms. The purpose of the study is to measure the performance of the ML model (KNN, RF, CatBoost, and the addition of data normalization pre-processing process) using the RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error) matrices.

2. RESEARCH METHODS

Research conducted by Fan Li dan Guang Jin [11] shows that the electricity load forecasting method using the designed KNN algorithm can analyze and forecast electricity loads in a short time with high forecasting accuracy. The research results also verify the validity and accuracy of the electricity load forecasting method. In the context of the continuous development of the electricity industry, electricity consumption technology in China is moving towards a more diverse, distributed, energy-saving, and intelligent direction.

Ergi Putra Febtiawan [17] conducted a study in the form of photovoltaic energy forecasting using the RF algorithm showing good results with high accuracy with a value of 96% and an error of 4% in predicting photovoltaic power system energy production. The results of forecasting photovoltaic power system energy production in the future show significant potential in supporting renewable energy needs, with consistent and reliable production estimates.

The research of Karthick Kanagarathinam and Ramasamy Dharmaprakash [26] used a dataset covering 11 attributes and 35,040 data. The CatBoost prediction algorithm was used to predict energy consumption and hyperparameter optimization using GridSearchCV with 5-fold cross-validation. The proposed model successfully predicted energy consumption for various types of loads with impressive results on both the training dataset (RMSE=0.382, R^2=0.999, MAPE=1.139) and the test dataset (RMSE=1.073, R^2=0.998, MAPE=1.142). These findings highlight the potential of CatBoost as a valuable tool for energy management and conservation, enabling organizations to make better decisions, optimize resource allocation, and support sustainability.

Based on previous research, this study compares ML methods, namely the Random Forest (RF), k-Nearest Neighbors (KNN), and CatBoost algorithms. The prediction results are measured using RMSE and MAPE matrices. The research diagram is presented in Figure 1.

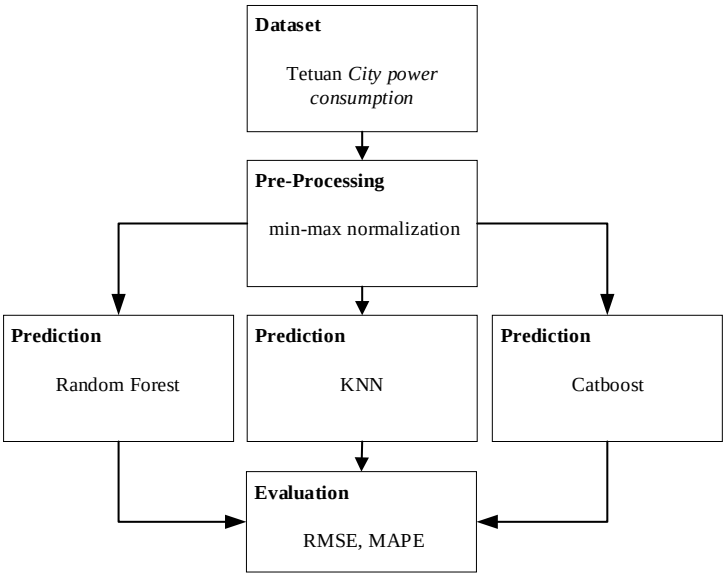

Figure 1. Research Diagram

The research began with the search for electricity consumption data. The dataset was taken from Kaggle on the electricity consumption data of Tetouan City. The Tetouan City dataset is multivariate, allowing a better

modeling approach in capturing the complex relationships between various factors that influence energy use. The dataset obtained was then subjected to a pre-processing stage in the form of min-max normalization, where the data was made in the range of 0 and 1. The results of the min-max normalization were then divided into training data and test data. The proportion of training data and test data is 90%:10%, 80%:20%, and 50%:50%. Consumption prediction uses three ML methods, namely RF, k-NN, and Catboost. The results of the ML method are measured based on the RMSE and MAPE values.

2.1 Pre-Processing

The pre-processing stage is carried out so that the data can be processed to the next stage. In this process, the data normalization process is carried out using the min-max normalization method. Equation (1) is the min-max calculation.

$$x' = \frac{(x\text{-}x_{min})}{(x_{max}\text{-}x_{min})} \quad (1)$$

where x is the original data value, x_{min} is the minimum feature data value and x_{max} is the highest feature data value.

2.2 Random Forest Algorithm

The random forest (RF) algorithm is a development of the Classification and Regression Trees (CART) method, namely by applying the bootstrap aggregating (bagging) and random feature selection methods. In random forests, many trees are grown to form a forest, then analysis is carried out on the collection of trees. How random forests work:

a. Perform the bootstrap stage, which randomly draws data of size n with recovery on the data cluster.
b. Using the bootstrap example, the tree is built until it reaches the maximum size (without pruning). At each node, the selection of the classifier is carried out by randomly selecting m explanatory variables, where m << p. The best classifier is selected from the m explanatory variables. This stage is the random feature selection stage.
c. Repeat steps a and b k times, until a forest consisting of k trees is formed.

2.3 KNN Algorithm

The prediction process is based on the neighborhood value, the neighbor classification is used as the prediction value of the test sample. The distance of the neighbor data is calculated based on the following Euclidean Equation (2).

$$\text{dist}\,(x,y) = \sqrt{\sum_{i=1}^{n}\left(x_i\text{-}y_i\right)^2} \quad (2)$$

where dist (x,y) is the proximity distance between data x to data y. x_i is testing data (test data) to i, and y_i is training data (training data) to i. n is number of attributes 1 to n.

2.4 CatBoost Algorithm

One of the algorithms that implement gradient boosting is CatBoost or categorical boosting. Model ranking generally uses Loss Function Change (LFC). The LFC value search uses Equations (3) and (4) below.

$$F = \{f_1, f_2, f_3, \dots, f_n\} \quad (3)$$

$$P_i = \beta F_j \quad (4)$$

F is a set of input features, β is a numerical factor assigned to the input features, and P is a prediction of a particular step. P_i is the predicted value of the substituted numerical factor, β represents the numerical factor, and F_j is a particular feature selected from the given set of features. Calculating the predicted value is shown in Equation (5).

$$P_{i+1} = \beta_{i+1} F_j \quad (5)$$

P_{i+1} represents the predicted value When the numerical factor is changed, β_{i+1} represents the modified numerical factor.

2.5 Evaluation

The calculation of RMSE and MAPE values is used to determine how well the model predicts the actual value, identify potential bias, and measure the level of prediction error. RMSE measures how far the predicted value is from the actual value. Equation (6) is the RMSE.

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (6)$$

where y_i is the original value, $\hat{y}_i$ is the predicted value, and n is the number of data.

MAPE is an evaluation scalar used to measure the average percentage error between the predicted value and the actual value. The calculation of the MAPE value is obtained from the absolute average of the percentage error shown in Equation (7).

$$MAPE=\frac{1}{M}\sum_{t=1}^{M}\left|\frac{y_t-\hat{y}_t}{y_t}\right| X\ 100\% \quad (7)$$

where M is the amount of data, y_t is the actual result value, and $\hat{y}_t$ is the predicted result value.

3. RESULTS AND DISCUSSION

3.1 Dataset

The initial data as in Table 1 used in the research is Kaggle data in the form of electrical energy consumption. 52416 data with attributes DateTime, Temperature, Humidity, Wind Speed, general diffuse flow, diffuse flow, Zone 1 Power Consumption, Zone 2 Power Consumption, and Zone 3 Power Consumption.

Table 1. Preliminary Data

Date Time	Tempe-rature	Humi-dity	Wind Speed	General Diffuse Flows	Diffuse Flows	Zone 1 Power Consumption	Zone 2 Power Consumption	Zone 3 Power Consumption
01/01/2017 00:00	6.559	73.8	0.083	0.051	0.119	34055.7	16128.88	20240.96
01/01/2017 00:10	6.414	74.5	0.083	0.07	0.085	29814.68	19375.08	20131.08
01/01/2017 00:20	6.313	74.5	0.08	0.062	0.1	29128.1	19006.69	19668.43
01/01/2017 00:30	6.121	75	0.083	0.091	0.096	28228.86	18361.09	18899.28
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
12/30/2017 23:30	6.9	72.8	0.086	0.084	0.074	29590.87	25277.69	13806.48
12/30/2017 23:40	6.758	73	0.08	0.066	0.089	28958.17	24692.24	13512.61
12/30/2017 23:50	6.58	74.1	0.081	0.062	0.111	28349.81	24055.23	13345.5

3.2 Pre-Processing

The technique used in the pre-processing process is min-max normalization. This process ensures that the data is on a scale between 0 and 1, making it easier to compare features that have different units. The min-max normalization process is carried out by reducing the minimum value of each feature from each data, then the results of the reduction are divided by the difference between the maximum and minimum values of the feature. Table 2 is the result of the min-max data normalization processing process.

Table 2. Data Normalization Value

Temperature	Humidity	Wind Speed	General Diffuse Flows	Diffuse Flows	Zone 1 Power Consumption	Zone 2 Power Consumption	Zone 3 Power Consumption
0.090091	0.748382	0.00513	4.04E-05	0.000115	0.526251	0.262361	0.343368
0.086146	0.75677	0.00513	5.67E-05	7.91E-05	0.415545	0.374886	0.340731
0.083399	0.75677	0.004663	4.99E-05	9.51E-05	0.397623	0.362116	0.329626
0.078176	0.762761	0.00513	7.48E-05	9.08E-05	0.374149	0.339738	0.311165
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
0.099366	0.736401	0.005596	6.88E-05	6.73E-05	0.409703	0.579491	0.188927
0.095504	0.738797	0.004663	5.33E-05	8.33E-05	0.393187	0.559197	0.181874
0.090662	0.751977	0.004819	4.99E-05	0.000107	0.377306	0.537116	0.177863

3.3 Prediction Process

The prediction process is carried out using the KNN, RF, and CatBoost algorithms. The normalized dataset is divided into two data, namely training data and testing data. The prediction process uses a comparison of training and testing data of 90%:10%, 80%:20%, and 50%:50%. Table 3 is a comparison of the number of training and testing data based on the total data.

Table 3. Comparison of Training and Testing Data

Criteria	Data Amount	
	Training	Testing
A= 90%:10%	47174	5242
B= 80%:20%	41932	10484
C= 50:50%	26208	26208

3.1.1. Prediction Using K-Nearest Neighbors (KNN) Algorithm

The KNN algorithm predicts new data based on the similarity to the closest training data. The working principle is to find the K training data closest to the new data, then the class or value of the K closest data is used as a prediction for the new data. The prediction results using data comparisons of 90%:10%, 80%:20%, and 50%:50% are presented in Table 4.

Table 4. Comparison of RMSE and MAPE Values in the KNN Algorithm

Value	A	B	C
RMSE	3213.9838	**3198.5983**	3261.7403
MAPE	0.1323	**0.1317**	0.1348

Table 4 shows that the comparison of training and testing data 80%:20% shows the best results in prediction based on the RMSE value of 3198.5983 and MAPE of 0.1317. Prediction using KNN and the pre-processing process of min-max data normalization obtained significant results. This can be seen in the results of the RMSE and MAPE values in Table 5. The prediction results of the KNN algorithm with the min-max data normalization process based on Table 5, a comparison of 90%:10% produces the smallest RMSE and MAPE values, namely 1314.1113 and 0.0458.

Table 5. Data Comparison of RMSE and MAPE Values in KNN Algorithm and Data Normalization

Value	A	B	C
RMSE	**1314.1113**	1427.5312	1649.4601
MAPE	**0.0458**	0.0493	0.0614

3.1.2. Prediction Using RF Algorithm

The experiment used the parameter of the number of decision trees to be built as many as 100 decision trees. The determination of weight initialization and data distribution is always the same every time the program is run using the random_state=42 command. The comparison of training and testing data is 90%:10%, 80%:20%, and 50%:50%. The prediction results based on the RMSE and MAPE values in the RF algorithm are presented in Table 6.

Table 6. Comparison of RMSE and MAPE Values in the RF Algorithm

Value	A	B	C
RMSE	**1335.4624**	1383.5631	1531.6175
MAPE	**0.0494**	0.0507	0.0580

The results of the smallest RMSE and MAPE values in the 90%:10% data comparison. The RMSE value is 1335.4624 and MAPE is 0.0494. The experimental results of the addition of the data normalization process to the RF algorithm are presented in Table 7.

The results of the smallest RMSE and MAPE values in the 90%:10% data comparison. The RMSE value is 1335.4624 and MAPE is 0.0494. The experimental results of the addition of the data normalization process to the RF algorithm are presented in Table 7.

Table 7. Comparison of RMSE and MAPE Values in RF Algorithm and Data Normalization

Value	A	B	C
RMSE	**1334.6781**	1383.7094	1531.8293
MAPE	**0.0493**	0.0507	0.0580

The experimental results obtained an insignificant value between the RF algorithm without min-max normalization compared to using Min-max normalization. This can be seen from the results of the RMSE value which only decreased by 0.7843 from the RMSE of the RF algorithm of 1335.46 to the RF algorithm with min-max normalization getting an RMSE of 1334.68while the MAPE value decreased by 0.0001. The change in value is not significant because RF in the selection of attributes for separation only depends on the order and distribution of values, not the scale of the data. Thus, changes in scale (such as normalization or standardization) do not affect the RF model.

3.1.3. Prediction Using CatBoost Algorithm

The CatBoost algorithm experiment used the parameters iterations=1000, learning_rate=0.1, depth=6, verbose=0. The model was built with 1000 decision trees. The more trees, the better the model is able to capture complex patterns in the data. A smaller learning rate (e.g. 0.01) makes the model learn slower and produces a more stable model, but requires more iterations. The depth=6 value indicates that the maximum depth of each tree is 6 levels. The results of the CatBoost algorithm experiment without min-max normalization using data comparisons of 90%:10%, 80%:20%, and 50%:50% get the results in Table 8

Table 8 Comparison of RMSE and MAPE in Algorithm CatBoost

Value	A	B	C
RMSE	**16343.157**	1658.1344	1682.7257
MAPE	**0.0717**	0.0719	0.0726

The experimental results show that the 90%:10% comparison produces the smallest RMSE and MAPE values, namely 1634.3157 and 0.0717. The prediction results using the addition of the min-max normalization process to the CatBoost algorithm are presented in Table 9.

Table 9. Comparison of RMSE and MAPE Values in the CatBoost Algorithm

Value	A	B	C
RMSE	**1634.4019**	1658.1343	1682.7250
MAPE	**0.0717**	0.0719	0.0726

The smallest RMSE and MAPE values are obtained from a 90%:10% data comparison. When compared to the results without the addition of min-max normalization, it is not significant, only 0.0862 for RMSE. This is because decision trees are the basis of its learning process, Catboost does not depend on the scale of the data.

3.4 Evaluation

The evaluation process is the final stage that aims to determine the performance of each algorithm in making predictions. Algorithm performance is measured using RMSE and MAPE. Table 10 is a comparison of algorithm performance without preprocessing with min-max normalization and using normalization.

Table 10. Comparison of RMSE and MAPE Values Based on 90% Training Data and 10% Testing Data

Dataset	Algorithm	Value	
		RMSE	MAPE
Original Data	KNN	3213.9838	0.1323
	RF	1335.4624	0.0494
	CatBoost	1634.3157	0.0717
Normalized Data	**KNN**	**1314.1113**	**0.0458**
	RF	1334.6781	0.0493
	CatBoost	1634.4019	0.0717

The prediction process using data normalization, the KNN algorithm is superior compared to the RF and CatBoost algorithms. The KNN value with data normalization gets an RMSE value of 1314.1113 and a MAPE of 0.0458. Table 10 shows that the results of RF and CatBoost do not change significantly when data normalization is added, because this algorithm only looks at the relative order or relative position of the data when dividing nodes in the tree, so the final result does not change with or without normalization. This is different from the KNN algorithm which works based on distance, which is very sensitive to data range. This was also stated by Dilber Uzun Ozsahin et al [27] that not all models require feature scaling techniques to be applied to the dataset to achieve optimal performance.

4. CONCLUSION

The performance results of each algorithm show that the method used is able to predict electricity consumption very well, this can be seen from the RMSE and MAPE values. The RMSE value is a measure of the average difference between the predicted value and the actual value. A smaller RMSE value indicates that the prediction model is more accurate in predicting the actual value. MAPE is a measure of the average percentage of absolute error between the predicted value and the actual value. A smaller MAPE value also indicates good algorithm performance in prediction. The performance of the KNN algorithm without using normalization gets an RMSE value of 3213.9838 and a MAPE value of 0.1323. After the preprocessing process is added, the RMSE value is 1314.1113 and the MAPE value is 0.0458. This shows that adding the min-max normalization process to the preprocessing process can improve the performance of the KNN algorithm in making predictions. The performance of the RF algorithm without normalization, the RMSE value is 1335.4626 and the MAPE value is 0.0494. Normalization produces constant RMSE and MAPE results, namely RMSE 1334.6781 and MAPE 0.0493. The performance of the CatBoost algorithm without using data normalization produces RMSE and MAPE values. Min-max normalization produces RMSE values of 1634.3157 and MAPE 0.0717. The RMSE of data normalization is 1634.4019 and MAPE is 0.0717. The RF and CatBoost algorithms perform better in handling data with different scales compared to KNN. In this study, KNN using data normalization outperforms RF and CatBoost. These findings highlight the important role of preprocessing in improving distance-based algorithms for energy consumption prediction, which can help in better energy management strategies. Further research is needed to evaluate the model on other datasets to ensure its generalization.

REFERENCE

[1] H. Hasanah and N. Nurmalitasari, "Aplikasi Sistem Cerdas Untuk Prediksi Energi Listrik Pemakaian Sendiri Di PT. Indonesia Power Sub Unit PLTA Kabupaten Wonogiri," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 8, no. 2, p. 489, 2017. https://doi.org/10.24176/simet.v8i2.1324

[2] D. Puspita, "Energi Bersih Dan Terjangkau Dalam Mewujudkan Tujuan Pembangunan Berkelanjutan (SDGs)," *J. Sos. dan sains*, vol. 4, no. 3, pp. 271–280, 2024, doi: https://doi.org/10.59188/jurnalsosains.v4i3.1245.

[3] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour, "Energy Forecasting: A Review and Outlook," *IEEE Open Access J. Power Energy*, vol. 7, pp. 376–388, 2020. https://doi.org/10.1109/OAJPE.2020.3029979

[4] A. Ibrahim, M. M. Muhammed, S. O. Sowole, R. Raheem, and O. Rabiat, "Performance of CatBoost classifier and other machine learning methods," *Data Sci.*, pp. 1–14, 2020, [Online]. Available: https://www.datasciencehub.net/system/files/ds-paper-644.pdf

[5] V. Kumar, N. Kedam, K. V. Sharma, D. J. Mehta, and T. Caloiero, "Advanced Machine Learning Techniques to Improve Hydrological Prediction: A Comparative Analysis of Streamflow Prediction Models," *Water (Switzerland)*, vol. 15, no. 14, 2023. https://doi.org/10.3390/w15142572.

[6] R. A. Asmara, Arief Prasetyo, Siska Stevani, and R. I. Hapsari, "Prediksi Banjir Lahar Dingin pada Lereng Merapi menggunakan Data Curah Hujan dari Satelit," *J. Inform. Polinema*, vol. 7, no. 2, pp. 35–42, 2021. https://doi.org/10.33795/jip.v7i2.494

[7] S. Widaningsih, "Penerapan Data Mining untuk Memprediksi Siswa Berprestasi dengan Menggunakan

Algoritma K Nearest Neighbor," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 3, pp. 2598–2611, 2022. https://doi.org/10.35957/jatisi.v9i3.859

[8] V. Ariyani, P. Putri, A. B. Prasetijo, and D. Eridani, "Perbandingan Kinerja Algoritme Naïve Bayes Dan K-Nearest Neighbor (Knn) Untuk Prediksi Harga Rumah," *J. Ilm. Tek. Elektro*, vol. 24, no. 2, pp. 162–171, 2022. https://doi.org/10.14710/transmisi.24.4.162-171

[9] P. Syahputra, "Prediksi Lama Rawat Pasien Covid-19 Berbasis Machine Learning," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 4, pp. 3374–3382, 2022. https://doi.org/10.35957/jatisi.v9i4.2883

[10] M. Yunus and N. K. A. Pratiwi, "Prediksi Status Gizi Balita Dengan Algoritma K-Nearest Neighbor (KNN) di Puskemas Cakranegara," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 4, no. 4, pp. 221–231, 2023. https://doi.org/10.35746/jtim.v4i4.328

[11] F. Li and G. Jin, "Research on power energy load forecasting method based on KNN," *Int. J. Ambient Energy*, vol. 43, no. 1, pp. 946–951, 2022. https://doi.org/10.1080/01430750.2019.1682041

[12] D. Eko Waluyo *et al.*, "Implementasi Algoritma Regresi pada Machine Learning untuk Prediksi Indeks Harga Saham Gabungan," *J. Inform. J. Pengemb. IT*, vol. 9, no. 1, pp. 12–17, 2024.

[13] O. H. Kombo, S. Kumaran, Y. H. Sheikh, A. Bovim, and K. Jayavel, "Long-term groundwater level prediction model based on hybrid KNN-RF technique," *Hydrology*, vol. 7, no. 3, pp. 1–24, 2020. https://doi.org/10.3390/HYDROLOGY7030059

[14] S. Huang, M. Huang, and Y. Lyu, "An Improved KNN-Based Slope Stability Prediction Model," *Adv. Civ. Eng.*, vol. 2020, no. 1, p. 16, 2020. https://doi.org/10.1155/2020/8894109

[15] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," *Big Data Min. Anal.*, vol. 4, no. 2, pp. 116–123, 2021. https://doi.org/10.26599/BDMA.2020.9020016

[16] N. R. Prasad, N. R. Patel, and A. Danodia, "Crop yield prediction in cotton for regional level using random forest approach," *Spat. Inf. Res.*, vol. 29, no. 2, pp. 195–206, 2021. https://doi.org/10.1007/s41324-020-00346-6.

[17] E. P. Febtiawan, L. A. Syamsul, I. Akbar, and A. S. Rachman, "Forecasting Produksi Energi Photovoltaic Menggunakan Algoritma Random Forest Classification," *J. Inf. Syst. Res.*, vol. 5, no. 4, pp. 1053–1062, 2024. https://doi.org/10.47065/josh.v5i4.5514

[18] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 12343 LNCS, pp. 503–515, 2001, doi: https://doi.org/10.1023/A:1010933404324

[19] M. Gholizadeh, M. Jamei, I. Ahmadianfar, and R. Pourrajab, "Prediction of nanofluids viscosity using random forest (RF) approach," *Chemom. Intell. Lab. Syst.*, vol. 201, no. March, p. 104010, 2020. https://doi.org/10.1016/j.chemolab.2020.104010

[20] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," *arXiv*, pp. 1–7, 2018. https://doi.org/10.48550/arXiv.1810.11363

[21] F. Ahmed, M. Saleem, Z. Rajpoot, and A. Noor, "Intelligent Heart Disease Prediction Using CatBoost Empowered with XAI," *Int. J. Comput. Innov. Sci.*, vol. 4, no. December, pp. 8–13, 2023.

[22] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost Classifier with other Machine Learning Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 738–748, 2020. https://doi.org/10.14569/IJACSA.2020.0111190.

[23] D. T. Tran, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Bilinear Input Normalization for Neural Networks in Financial Forecasting," 2021. https://doi.org/10.48550/arXiv.2109.00983

[24] H. Gadde, "AI-Assisted Decision-Making in Database Normalization and Optimization Hemanth Gadde," *Int. J. Mach. Learn. Res. Cybersecurity Artif. Intell.*, vol. 11, no. 01, pp. 230–259, 2020.

[25] A. Pranolo, F. Usha, and A. Khansa, "Enhanced Multivariate Time Series Analysis Using LSTM: A Comparative Study of Min-Max and Z-Score Normalization Techniques," vol. 16, no. 2, pp. 210–220, 2024.

[26] K. Karthick, R. Dharmaprakash, and S. Sathya, "Predictive Modeling of Energy Consumption in the Steel Industry Using CatBoost Regression: A Data-Driven Approach for Sustainable Energy Management," *Int. J. Robot. Control Syst.*, vol. 4, no. 1, pp. 33–49, 2024. https://doi.org/10.31763/ijrcs.v4i1.1234

[27] D. U. Ozsahin, M. Taiwo Mustapha, A. S. Mubarak, Z. Said Ameen, and B. Uzun, "Impact of feature scaling on machine learning models for the diagnosis of diabetes," in *2022 International Conference on Artificial Intelligence in Everything (AIE)*, pp. 87–94, 2022. https://doi.org/10.1109/AIE57029.2022.00024