Original Article

Available online at
www.heca-analitika.com/ijds

Infolitika Journal of Data Science

Vol. 3, No. 1, 2025

Optimizing Energy Consumption Prediction Across the IMT-GT Region Through PCA-Based Modeling

Muhammad Farid [1], Teuku Muhammad Faiz Nuzullah [1], Zatul Aklya [1], Syifa Nazila [1], Muhammad Zia Ulhaq [1], Feby Apriliansyah [2] and Novi Reandy Sasmita [1,*]

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; m.farid21@mhs.usk.ac.id (M.F.); tm_faiz@mhs.usk.ac.id (T.M.F.N.); zatun_a21@mhs.usk.ac.id (Z.A.); syifa_n21@mhs.usk.ac.id (S.N.); m_zia@mhs.usk.ac.id (M.Z.U.); novireandys@usk.ac.id (N.R.S.)

[2] Department of Regional and Urban Planning, Faculty of Engineering, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; feby.aprilians@gmail.com (F.A.)

* Correspondence: novireandys@usk.ac.id

Article History

Received March 13 2025
Revised 13 May 2025
Accepted 22 May 2025
Available Online 28 May 2025

Keywords:
Principal Component Analysis
Random Forest
Energy consumption
Prediction
Multicollinearity

Abstract

This study aims to improve the accuracy of energy consumption prediction in the Indonesia-Malaysia-Thailand Growth Triangle (IMT-GT) region by addressing multicollinearity among independent variables such as energy production (Mtoe), lignite coal production (million tons), crude oil production (million tons), refined oil production (million tons), natural gas production (billion cubic meters), and electricity production (terawatt-hours). By integrating Principal Component Analysis (PCA) with Random Forest (RF), six correlated variables were reduced into two uncorrelated principal components (PC1 and PC2), explaining 80.77% of the data variance. The PCA-RF hybrid model outperformed the standalone Random Forest (RF) model, with an increase in the coefficient of determination (R^2) from 0.976 to 0.993. Additionally, it achieved significant reductions in error metrics, with the mean absolute error (MAE) decreasing from 5.811 to 4.169 and the root mean square error (RMSE) dropping from 9.278 to 4.786. These results demonstrate PCA's effectiveness in isolating dominant drivers such as energy and lignite coal production while improving model stability. The framework provides policymakers with a reliable tool to forecast energy demand and align economic growth with sustainability in fossil fuel-dependent economies.

1. Introduction

Cooperation between countries is important to achieving common goals in various fields, especially in supporting economic growth [1]. Through such cooperation, countries can help each other face challenges that are not easily resolved independently. As a form of regional synergy, Indonesia, Malaysia, and Thailand formed an economic cooperation known as the Indonesia-Malaysia-Thailand Growth Triangle (IMT-GT). IMT-GT cooperation has shown significant growth in recent years, especially in the investment, economic, and infrastructure development sectors. However, along with this growth, new challenges have emerged related to increased energy consumption in the three countries [2]. Economic growth and industrial expansion in the region have the potential to put pressure on environmental sustainability and a stable energy supply. Rapid economic growth and increased industrial activities may lead to changes in energy consumption [3]. Therefore, efforts should be made to optimize energy consumption predictions to more effectively anticipate future energy needs.

Increasing energy consumption in developing countries, such as Indonesia, Malaysia, and Thailand, generally goes hand in hand with industrialization and urbanization. Energy consumption in these regions is highly correlated with the rate of economic growth and industrial sector activity [4]. On the other hand, the high dependence on fossil energy, such as coal and petroleum, poses energy efficiency and sustainability challenges. An accurate and comprehensive data-driven analytical approach is needed to understand and map the complex dynamics between the various factors that influence energy consumption at the regional level.

The Indonesia-Malaysia-Thailand Growth Triangle (IMT-GT), established in 1993, represents a strategic economic collaboration aimed at accelerating regional development through cross-border industrialization, infrastructure expansion, and trade integration [5]. Over the past decade, the IMT-GT has witnessed unprecedented economic growth, with GDP contributions from the three nations rising by an average of 5.2% annually [6]. However, this rapid industrialization and urbanization have intensified energy consumption, driven by sectors such as manufacturing, transportation, and utilities. For instance, Thailand's energy demand surged by 28% between 2015 and 2022, while Indonesia and Malaysia reported similar trends, with fossil fuels accounting for over 75% of their energy mix [7]. This dependency raises critical concerns about environmental sustainability, energy security, and the feasibility of achieving net-zero emissions targets by 2050 [8].

Accurate energy consumption prediction is pivotal for balancing economic growth with sustainable resource management. However, the IMT-GT's energy landscape is characterized by complex interdependencies among variables such as lignite coal production, natural gas output, and electricity generation, which exhibit strong multicollinearity [9]. Traditional prediction models and standalone machine learning algorithms, often fail to account for these collinear relationships, resulting in overfitting and reduced generalizability. Recent studies highlight that oversights in handling multicollinearity can inflate prediction errors by up to 30% in energy systems, underscoring the need for advanced methodologies tailored to the region's unique dynamics [10].

Machine learning (ML) models, particularly Random Forest (RF), have gained traction in energy forecasting due to their ability to capture non-linear relationships and handle heterogeneous datasets [11]. For example, Zhang et al. applied RF to predict urban energy demand in Southeast Asia, achieving a mean absolute error (MAE) of 8.2% [12]. Similarly, Joo et al. utilized RF to model industrial energy consumption in Vietnam, demonstrating its superiority over regression-based approaches. Despite these advancements, a critical limitation persists: RF's performance deteriorates when input variables are highly correlated, as it prioritizes redundant features during tree construction, thereby obscuring dominant drivers [13, 14].

The primary challenge in IMT-GT energy modeling lies in disentangling multicollinear variables to isolate their individual and collective impacts on consumption patterns. Conventional ML approaches, including standalone RF, inadequately address this complexity, leading to biased estimates and unreliable policy recommendations. To bridge this gap, dimensionality reduction techniques like Principal Component Analysis (PCA) offer a robust solution. PCA transforms correlated variables into orthogonal principal components (PCs), preserving >95% of data variance while eliminating redundancy. When integrated with RF, PCA enhances model interpretability and stability by distilling multicollinear datasets into uncorrelated inputs, thereby mitigating overfitting risks [15].

Recent studies demonstrate the efficacy of hybrid PCA-ML frameworks in energy forecasting. For instance, Sun et al. combined PCA with gradient-boosted trees to predict residential energy use in China, reducing prediction errors by 22% compared to baseline models [16]. Similarly, Testasecca et al. applied a hybrid model combining PCA with Long Short-Term Memory (LSTM) networks to forecast renewable energy generation in India, resulting in an improvement in forecasting accuracy. The application of PCA prior to LSTM helps reduce dimensionality while preserving relevant information, which is crucial for improving model performance in time-series forecasting [17]. These successes highlight PCA's versatility in enhancing ML models, yet its application remains unexplored in the IMT-GT context. This study adapts this hybrid approach to address the region's unique energy dynamics, where cross-border industrial synergies and fossil fuel dependencies amplify variable correlations.

This study aims to optimize the prediction of energy consumption in the IMT-GT region by integrating PCA with RF to overcome multicollinearity. It is hypothesized that PCA-based dimensionality reduction will significantly improve the accuracy of RF by isolating the dominant drivers. The novelty of this study introduces a novel integration of PCA with RF to explicitly address multicollinearity in energy consumption prediction - a critical gap in previous studies. Unlike conventional approaches that apply RF in isolation, the PCA-based framework transforms correlated variables into orthogonal principal components, thus eliminating redundancy while retaining important information. This approach not only advances methodological rigor, but also supports sustainable energy planning in developing countries.

2. Materials and Methods

2.1. Data Descriptions

The data used in this study is secondary data sourced from the *World Energy & Climate Statistics - Yearbook 2023* publication. This publication includes data on various energy and climate elements, such as consumption, production, emissions, efficiency, and energy technologies across countries and regions. This information can be used to analyze and compare global energy and climate trends and conditions and to develop policies and strategies aligned with sustainable development goals. The data used in this study focuses on energy consumption in the IMT-GT region, namely Indonesia, Malaysia, and Thailand. The data is analyzed using the PCA method to address the issue of multicollinearity among the independent variables.

This study uses seven variables, consisting of one dependent variable and six independent variables. The dependent variable is energy consumption, which is measured in million tons of oil equivalent (Mtoe). The independent variables include energy production (Mtoe), lignite coal production (million tons), crude oil production (million tons), refined oil production (million tons), natural gas production (billion cubic meters), and electricity production (terawatt-hours).

2.2. Data Preprocessing

2.2.1. Energy Consumption

Energy consumption is the amount of energy used or utilized by an activity, sector, or region within a certain period; energy consumption also impacts the environment, health, and welfare [18]. The role of energy is crucial to the advancement of society in an era of ever-evolving industry and technology. Human productivity in the workplace is increasing thanks to technological advancements, which drive economic growth [19]. Energy is essential for running various economic activities, such as meeting daily consumption needs and supporting various production processes. Energy management must be done responsibly based on the principle of sustainable development, as energy is an irreplaceable natural resource [20]. To maximize its benefits for society's overall prosperity, energy must be utilized to maintain sustainability and natural balance.

A country's economic development is highly dependent on energy consumption. The main reasons for the increase in energy demand are rapid population growth, changing lifestyles, increasing production sectors, and fierce economic competition [21]. In such a situation, energy sources such as oil, coal, and electricity are essential to fulfill energy needs and support the dynamics of a growing economy. Coal plays an important role in heavy industry, while oil is essential to transportation and mobility. In addition, electricity, which is increasingly turning to renewable sources, serves as the backbone for various sectors of the economy. Therefore, to foster sustainable economic growth, management and optimization of various energy resources is essential [22].

2.2.2. IMT-GT region

The Indonesia, Malaysia, and Thailand Growth Triangle (IMT-GT) is a regional economic cooperation program under ASEAN. This subregional economic cooperation program was established on July 20, 1993, at the Ministerial Meeting (PTM) held in Langkawi, Malaysia. The purpose of establishing IMT-GT is to accelerate the economic transformation of member countries and provinces in the three countries by complementing each other's basic needs and utilizing their comparative advantages. Since its establishment, IMT-GT has expanded its geographical scope to 32 provinces and three countries; 10 provinces in Indonesia, eight states in western Malaysia, and 14 Thailand have become members [23]. Due to the many connections, including geographical, historical, cultural, and linguistic links, the member countries and provinces are well suited to cooperate in the economy.

The seven key areas of IMT-GT cooperation include tourism, trade and investment, transportation, agriculture, environment, human resources, and halal cooperation [24]. In this area, increased cooperation directly impacts energy consumption and production. Measures to improve efficiency and sustainability in energy consumption are necessary to ensure that rapid economic development does not negatively impact the environment [25].

Increased economic activity in tourism, trade, and investment will increase energy consumption, including oil, coal, and electricity. Therefore, it is necessary to focus on diversifying energy sources towards more sustainable ones, such as using renewable energy, developing green technologies, and implementing policies that support energy efficiency. These steps are taken wisely to ensure economic growth provides economic benefits while contributing to environmental preservation in the IMT-GT region [26].

2.2.3. Random Forest (RF)

RF is a development of the Classification and Regression Tree (CART) method that applies random feature selection and bootstrap aggregating or bagging [27, 28]. The final class prediction in the RF algorithm uses the prediction results of each tree combined or aggregated. From the collection of predictions, the most votes are selected in the class category, which often appears as a prediction of the classification tree (majority voting). The classification results are influenced by the adjustment of model parameters commonly known as hyperparameters. Different hyperparameter values result in different model parameter values for a given data set.

2.2.4. Principal Component Analysis

PCA is an analytical method that was first discovered in 1901 by Karl Pearson. PCA is used to simplify data by transforming data into linear to produce a new coordinate system with maximum variance [29]. PCA is an analysis that aims to reduce the dimensions of data that are multicollinear to data dimensions that are not multicollinear without significantly reducing their characteristics [30].

Multicollinearity is a situation with a strong correlation or relationship between independent variables. Multicollinearity can occur when the correlation coefficient is close to one or greater than 0.75 [31]. The correlation coefficient can be calculated using the Pearson correlation formula, which is as shown in Equation 1:

$$r_{X_jX_l} = \frac{n\sum_{j=1}^{n}\sum_{l=1}^{n} X_j X_l - \sum_{j=1}^{n} X_j \sum_{l=1}^{n} X_l}{\sqrt{n\sum_{j=1}^{n} X_j^2 - \left(\sum_{j=1}^{n} X_j\right)^2}\sqrt{n\sum_{l=1}^{n} X_l^2 - (\sum_{l=1}^{n} X_l)^2}} \quad (1)$$

Where n refers to the number of observations, X_j and X_l are the j-th and the l-th variables, respectively, and k is the total number of variables.

Several strategies can be applied to address multicollinearity, including removing highly correlated independent variables, adding additional data to improve variability and reduce correlation, transforming the variables to alter their relationships, or applying PCA, which converts the original correlated variables into a new set of uncorrelated principal components [32]. This process involves several key stages: preparing and standardizing the data, computing the variance-covariance matrix, extracting eigenvalues and eigenvectors, determining the proportion of variance explained, and calculating principal component scores.

2.2.4.1. Data Preparation

There are two data preparation processes: data cleaning and data standardization. Data cleaning eliminates incomplete and accurate data to produce quality data [33]. Data standardization is done to ensure that each variable contributes equally to the analysis results [34]. Data standardization can be calculated using the Z-score method, as shown in Equation 2:

$$Z_{ij} = \frac{(X_{ij} - \mu_j)}{\sigma_j} \quad (2)$$

Where X_{ij} refers to the original value of the i-th observation for the j-th variable, μ_j is the mean of the j-th variable, and σ_j denotes the standard deviation associated with that variable. This transformation allows each variable to be centered around zero with a standard deviation of one, facilitating comparison across variables measured on different scales.

2.2.4.2. Variance-Covariance Matrix Calculation

Variance is calculated to assess the spread or dispersion of data in a study, while the covariance matrix represents the covariances between pairs of variables, with each cell in the matrix containing a covariance value derived from the sample data [35]. The variance is computed using the Equation 3:

$$Var(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Z_{ij} - \mu_j\right)^2 \quad (3)$$

This measures the average squared deviation of standardized values from their mean. Meanwhile, the covariance between two variables is calculated using Equation 4:

$$Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_{xj})(y_{ij} - \mu_{yj}) \quad (4)$$

Where μ_x and μ_y refers to sample means of variables x and y respectively and x_i and y_i *i-th* observation value of x and y variables

2.2.4.3. Eigenvalue and Eigenvector Calculation

To identify the principal components, calculate the eigenvalues (λ) and eigenvectors (v) of the variance-covariance matrix (R). The eigenvalues are determined by solving the characteristic equation, as shown in Equation 5:

$$det\ (R - \lambda I) = 0 \quad (5)$$

For each eigenvalue, the corresponding eigenvector satisfies the Equation 6:

$$Rv = \lambda v \tag{6}$$

2.2.4.4. Proportion of Principal Components

The proportion of variance explained by each principal component (PC) is calculated as shown in Equation 7:

$$PC(\%) = \frac{Eigenvalue}{Total\ Variance} \times 100\% \tag{7}$$

This value helps determine how much of the total dataset variability is captured by each component.

2.2.4.5. Principal Component Scores

Once significant principal components are selected (typically those with eigenvalues greater than one), calculate the principal component scores for each observation. This is done by projecting the standardized data onto the eigenvectors, as shown in Equation 8:

$$A_j = v_j' Z_j \tag{8}$$

Where $v_j'$is the transpose of the j-th eigenvector of the covariance matrix S, and Z is the vector of standardized variables. This transformation is a key step in PCA, which aims to reduce a dataset's dimensionality while retaining most of the original variability.

2.2.4.6 Selecting the Number of Principal Components

The number of principal components (PC) can be determined in three ways: principal components with more than one eigenvalue, principal components with a total variance value obtained of more than 80%, and the number of major components by looking at the fracture at the elbow of the scree plot [36]. However, some experts recommend choosing principal components with an eigenvalue greater than one. This is because the explained diversity of the data will be reduced if the eigenvalue is less than one.

3. Results and Discussion

3.1. Descriptive Statistics

Table 1 summarizes descriptive statistics of the variables related to energy production and consumption in the IMT-GT cooperation region, including the minimum value, first quartile (Q1), median, mean value, third quartile (Q3), and maximum value. These variables include energy production, lignite coal production, crude oil production, refined oil production, natural gas production, electricity production, and energy consumption. The descriptive analysis provides an overview of the central tendency and dispersion of the data, helping to identify patterns, outliers, and potential data inconsistencies. For instance, variables with large gaps between the minimum and maximum values may indicate significant variability among countries or time periods.

Table 1 shows that Energy Production has the highest average of 150.4 and a wide range of values, from 27 to 470, indicating high variability. Similarly, Lignite Coal Production has a minimum value of 0 and a maximum of 601, with an average of 89.88 but a median of only 18, indicating a highly skewed distribution to the right. In contrast, Crude Oil and Refined Oil Production shows a more balanced distribution, indicated by relatively close mean and median values and a narrow interquartile range, signaling the stability of production in the category.

Meanwhile, Natural Gas Production has a median (56) slightly higher than the mean (48.74), indicating a mild negative distribution trend. Electricity Production displays a high average (124.43) with a wide range, indicating large fluctuations in electricity production. On the consumption side, Energy Consumption has a mean of 110.8 and a median of 99.5, with a third quartile of 139.2, indicating that the energy consumption pattern is also quite varied. In comparison, energy production is generally higher than consumption, which may indicate potential energy exports or inefficiencies in energy distribution and utilization.

3.2. Multicollinearity Detection

Multicollinearity testing uses Pearson correlation, where if the correlation value between independent variables is more than 0.75, a multicollinearity problem in the data can be assumed. The following is a correlation matrix or covariance variance matrix shown in Figure 1.

The matrix shows a high correlation between the energy production variable and lignite coal production, which is 0.92. So, it can be concluded that there is a multicollinearity problem in this study data because the Pearson correlation value between several independent variables is more than 0.75. So, it is necessary to take further action to overcome the multicollinearity problem in this study data using PCA.

3.3 Principal Component Analysis (PCA)

Before conducting the PCA, all independent variables were standardized to ensure comparability, as they were originally measured on different scales. The standardized variables are denoted as follows: Z1 represents Energy Production, Z2 represents Lignite Coal Production, Z3 represents Crude Oil Production, Z4 represents Refined Oil Production, Z5 represents Natural Gas Production, and Z6 represents Electricity Production.

Table 1. Descriptive statistics of energy-related variables in the IMT-GT region.

Variables	Min	Q1	Median	Mean	Q3	Max
Energy Production	27	64	88	150.4	228.2	470
Lignite Coal Production	0	2.75	18	89.88	56.5	601
Crude Oil Production	3	17.75	33	33.94	41.25	81
Refined Oil Production	10	24	42	37.54	49	65
Natural Gas Production	6	27	56	48.74	70	86
Electricity Production	23	73.25	118.5	124.43	170	309
Energy Consumption	21	69	99.5	110.8	139.2	241

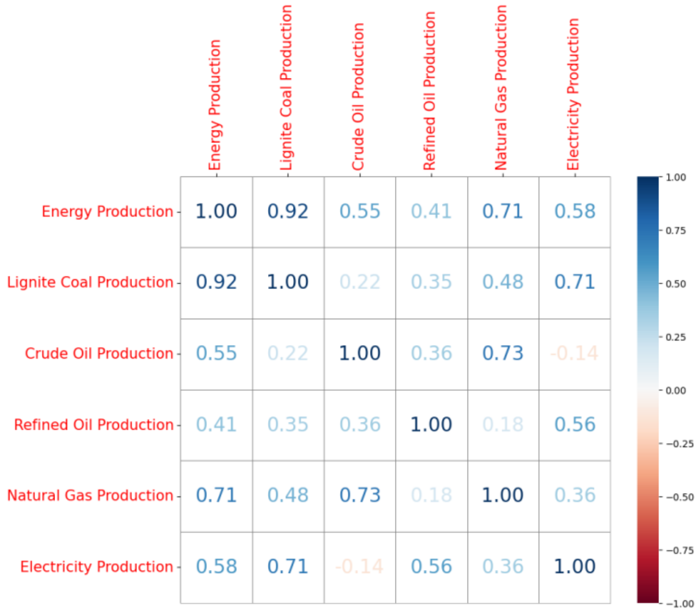

Figure 1. Correlation matrix of independent variables.

Table 2. Results of principal component analysis.

Variables	Key Components					
	PC1	PC2	PC3	PC4	PC5	PC6
Z1	0.5232	0.0636	0.1761	-0.3382	-0.1818	0.7374
Z2	0.4735	-0.1962	0.4016	-0.4249	-0.0638	-0.6256
Z3	0.2918	0.6511	-0.2915	-0.1645	0.6037	-0.1202
Z4	0.3026	-0.3337	-0.8435	-0.1208	-0.2466	-0.1007
Z5	0.4206	0.3994	0.0553	0.6471	-0.4637	-0.1636
Z6	0.3853	-0.5125	0.0902	0.4946	0.568	0.1162
Eigenvalue	3.3841	1.4619	0.7289	0.3679	0.0551	0.0021
Variance Proportion	0.5640	0.2436	0.1215	0.0613	0.0092	0.0004
Cumulative Proportion of Variance	0.5640	0.8077	0.9292	0.9905	0.9996	1

In this study, the Kaiser criterion, which considers eigenvalues that are more than one, is used to determine the number of main components. Based on Table 2, only two principal components, PC1 and PC2, meet these criteria. This shows that the first two principal components are sufficient to represent most of the information from the original data.

Table 2 shows two main components whose eigenvalue is more than one; PC1 has an eigenvalue of 3.3841 with a proportion of variance of 56.40%, while PC2 has an eigenvalue of 1.4619 with a proportion of variance of 24.36%. The two main components have a cumulative variance proportion of 80.77%; this means that PC1 and PC2 can explain 80.77% of the data diversity of the six independent variables. The equation obtained to calculate the score of the selected main component are shown in Equation 9 and Equation 10:

$$\begin{aligned} PC1 = 0.5232Z1 + 0.4735Z2 + 0.2918Z3 \\ + 0.3026Z4 + 0.4206Z5 \\ + 0.3853Z6 \end{aligned} \tag{9}$$

$$\begin{aligned} PC2 = 0.0636Z1 - 0.1962Z2 + 0.6511Z3 \\ - 0.3337Z4 + 0.3994Z5 \\ - 0.5125Z6 \end{aligned} \tag{10}$$

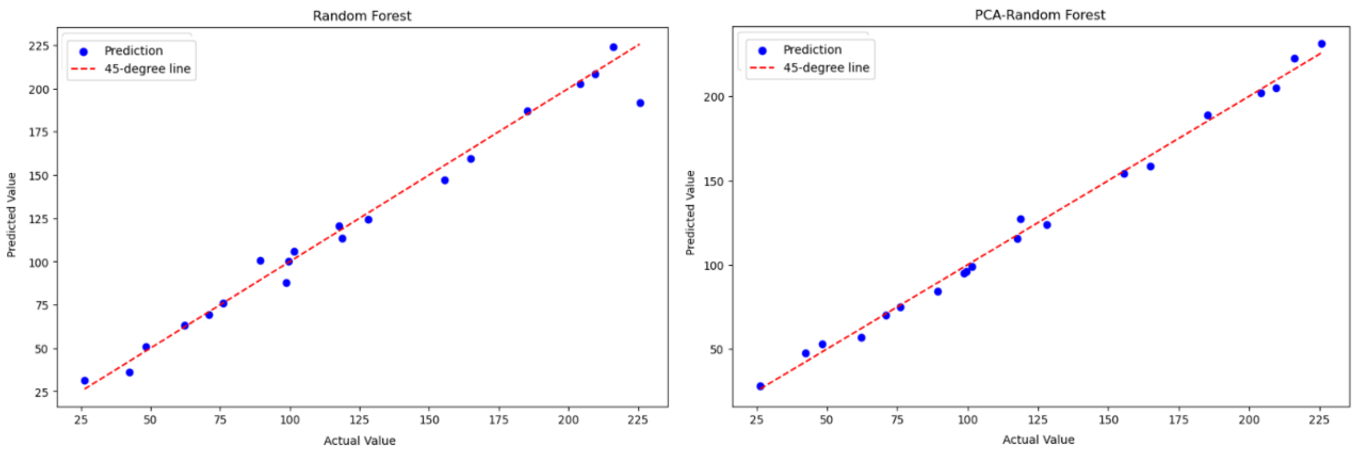

Figure 2. Comparison of actual vs. predicted values for RF and PCA-RF models.

Table 3. Comparison of RF and PCA-RF model evaluation.

Evaluation Metrics	RF	PCA-RF
MAE	5.811	4.169
MSE	86.091	22.909
RMSE	9.278	4.786
R^2	0.976	0.993

3.4. Comparison of RF and PCA-RF Predictions

Table 3 presents the results of evaluating model performance using several error metrics commonly used in regression analysis: MAE, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). These four metrics measure the error rate and prediction accuracy of the two models being compared: the RF model and the PCA-RF model.

Table 3 shows that the PCA-RF model has a lower MAE value than the RF model, which is 4.169 compared to 5.811. The MSE value also decreases from 86.091 to 22.909, and the RMSE value drops from 9.278 to 4.786. A high R^2 value indicates that the model has a high level of accuracy. Table 3 shows that the PCA-RF model has a higher R^2 value than the RF model, which is 0.993 compared to 0.976. This result shows that the PCA-RF model has better prediction performance and is more accurate than the RF model without the application of PCA.

Figure 2 presents a visualization of the prediction results against the actual values for the two models compared in this study, namely RF and PCA-RF. Each graph shows the scatter between the actual and predicted values produced by the models, with a 45-degree line as a reference to show perfect agreement between the two. This visualization aims to provide further insight into the predictive performance of the two models while supporting the quantitative analysis described in Table 3.

Based on the scatterplot above, the RF model without applying PCA has shown quite good performance. However, after PCA is applied before the RF algorithm, the feature dimensions are effectively reduced to minimize the model complexity and potential multicollinearity. As a result, the model becomes simpler and more accurate than before. Visually, the prediction points on the scatterplot after the application of PCA look closer to the line, which indicates that the model predictions are getting closer to the actual values. This shows that applying PCA to the RF algorithm can improve the model's ability to capture intrinsic patterns in the data more efficiently and accurately.

4. Conclusions

Based on the results of the study, it can be concluded that there is a multicollinearity problem in energy consumption data in the IMT- GT cooperation area. To overcome this problem, the PCA method is used, which successfully reduces the data dimension into two main components (PC1 and PC2), which are cumulatively able to explain 80.77% of the data diversity. The model evaluation results show that the application of PCA can significantly improve the performance of the RF model. This is shown by the decrease in prediction error values: MAE from 5.811 to 4.169, MSE from 86.091 to 22.909, and RMSE from 9.278 to 4.786. In addition, the coefficient of determination (R^2) increased from 0.976 to 0.993, indicating that the model with PCA has a more accurate prediction ability and can better explain the variability of energy consumption data than the model without PCA. Thus, the application of PCA in building an energy consumption prediction model proved effective in improving the accuracy and stability of the model.

Author Contributions: Author Contributions: Conceptualization, Z.A., M.F., T.M.F.N., and S.N.; methodology, M.F., N.R.S., Z.A., and S.N.; software, M.F., Z.A., and N.R.S.;

validation, N.R.S., S.N., and M.Z.U.; formal analysis, Z.A., M.F., and S.N.; investigation, M.Z.U., N.R.S., and S.N.; resources, N.R.S., T.M.F.N., and Z.A.; data curation, N.R.S., Z.A., and S.N.; writing—original draft preparation, Z.A., M.F., F.A., and S.N.; writing—review and editing, N.R.S., M.Z.U, F.A, and Z.A.; visualization, T.M.F.N., and M.Z.U.; supervision, N.R.S.; project administration, N.R.S. All authors have read and agreed to the published version of the manuscript

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available upon request to the authors.

Acknowledgments: Thanks to World Energy & Climate Statistics - yearbook 2023 publication for providing free access to data and making valuable contributions supporting this study.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Hardi, I., Ray, S., Attari, M. U. Q., Ali, N., and Idroes, G. M. (2024). Innovation and Economic Growth in the Top Five Southeast Asian Economies: A Decomposition Analysis, *Ekonomikalia Journal of Economics*, Vol. 2, No. 1, 1–14. doi:10.60084/eje.v2i1.145.
2. Zhang, M., Zuo, Y., Yu, X., Zhang, A., and Zhou, H. (2022). Experimental Study on Seepage Flow Characteristics of Leaked Molten Salt in Foundation Fillers in Concentrating Solar Power Plants, *Asia-Pacific Journal of Chemical Engineering*, Vol. 17, No. 5, e2813. doi:10.1002/apj.2813.
3. Ajaz Ahmed. (2022). Energy Smart Buildings: Potential for Conservation and Efficiency of Energy, *The Pakistan Development Review*, Vol. 53, No. 4, 371–381. doi:10.30541/v53i4IIpp.371-381.
4. Sanusi, A., Singagerda, F. S., and Sani, A. Z. (2021). World Oil Price Shocks in Macroeconomic ASEAN +3 Countries: Measurement of Risk Management and Decision-Making a Linear Dynamic Panel Approach, *International Journal of Energy Economics and Policy*, Vol. 11, No. 4, 75–83. doi:10.32479/ijeep.11249.
5. Mahdi, M. A., Hosny, K. M., and Elhenawy, I. (2021). Scalable Clustering Algorithms for Big Data: A Review, *IEEE Access*, Vol. 9, No. 1, 80015–80027. doi:10.1109/ACCESS.2021.3084057.
6. World Bank Expects Solid Growth but Risky Outlook for South Asia. (2023, March).
7. ASEAN Centre for Energy (ACE), Kresnawan, Muhammad, R., and Beni, S. (2021). ASEAN Energy in 2022: Outlook Report, *ASEAN Centre for Energy*, 1–26.
8. Dawson, B., and Spannagle, M. (2020). *Intergovernmental Panel on Climate Change (Ipcc), The Complete Guide to Climate Change*. doi:10.4324/9780203888469-41.
9. Wang, S., Lu, H., Khan, A., Hajati, F., Khushi, M., and Uddin, S. (2022). A Machine Learning Software Tool for Multiclass Classification, *Software Impacts*, Vol. 13, No. 1, 100383. doi:10.1016/j.simpa.2022.100383.
10. Olayemi, Olanrewaju, S. (2020). Effects of Multicollinearity and Correlation between the Error Terms on Some Estimators in a System of Regression Equations, *Global Journal of Science Frontier Research*, Vol. 1, No. 1, 77–94. doi:10.34257/GJSFRFVOL20IS4PG77.
11. Zubair, M., Iqbal, M. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., and Sarker, I. H. (2024). An Improved K-Means Clustering Algorithm towards an Efficient Data-Driven Modeling, *Annals of Data Science*, Vol. 11, No. 5, 1525–1544. doi:10.1007/s40745-022-00428-2.
12. Zhang, S., Wang, H., and Keller, A. A. (2021). Novel Machine Learning-Based Energy Consumption Model of Wastewater Treatment Plants, *ACS ES&T Water*, Vol. 1, No. 12, 2531–2540. doi:10.1021/acsestwater.1c00283.
13. Joo, C., Park, H., Lim, J., Cho, H., and Kim, J. (2022). Development of Physical Property Prediction Models for Polypropylene Composites with Optimizing Random Forest Hyperparameters, *International Journal of Intelligent Systems*, Vol. 37, No. 6, 3625–3653. doi:10.1002/int.22700.
14. Maulana, A., Idroes, G. M., Kemala, P., Maulydia, N. B., Sasmita, N. R., Tallei, T. E., Sofyan, H., and Rusyana, A. (2023). Leveraging Artificial Intelligence to Predict Student Performance: A Comparative Machine Learning Approach, *Journal of Educational Management and Learning*, Vol. 1, No. 2, 64–70. doi:10.60084/jeml.v1i2.132.
15. Rodionova, O., Kucheryavskiy, S., and Pomerantsev, A. (2021). Efficient Tools for Principal Component Analysis of Complex Data— a Tutorial, *Chemometrics and Intelligent Laboratory Systems*, Vol. 213, No. 1, 104304. doi:10.1016/j.chemolab.2021.104304.
16. Sun, Z., Fan, H., Yuan, Q., Liu, S., Wang, Y., Li, X., Zhang, Y., and Wang, Y. (2024). Area Control Error Prediction Based on PCA-XGBoost. doi:10.3233/FAIA231317.
17. Testasecca, T., Lazzaro, M., Sarmas, E., and Stamatopoulos, S. (2023). Recent Advances on Data-Driven Services for Smart Energy Systems Optimization and pro-Active Management, *2023 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv)*, IEEE, 146–151. doi:10.1109/MetroLivEnv56897.2023.10164056.
18. Koengkan, M., Fuinhas, J. A., Kazemzadeh, E., Osmani, F., Alavijeh, N. K., Auza, A., and Teixeira, M. (2022). Measuring the Economic Efficiency Performance in Latin American and Caribbean Countries: An Empirical Evidence from Stochastic Production Frontier and Data Envelopment Analysis, *International Economics*, Vol. 169, 43–54. doi:10.1016/J.INTECO.2021.11.004.
19. Balgabekova, Z., Alaran, M., and Varol, H. A. (2023). A Data-Centric Approach for Object Recognition in Hemispherical Camera Images, *IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society*, IEEE, 1–6. doi:10.1109/IECON51785.2023.10311869.
20. Zhou, J., Xu, L., Zhang, L., Zhang, J., and Zhang, A. (2024). Electrospun Polyvinylidene Fluoride Nanogenerator with Hydrophobic Properties and Enhanced Energy Harvesting for Wearable and Wireless Sensing, *ACS Applied Electronic Materials*, Vol. 6, No. 3, 2026–2038. doi:10.1021/acsaelm.4c00053.
21. Awosusi, A. A., Rjoub, H., Ağa, M., and Onyenegecha, I. P. (2024). An Insight into the Asymmetric Effect of Economic Globalization on Renewable Energy in Australia: Evidence from the Nonlinear ARDL Approach and Wavelet Coherence, *Energy & Environment*, Vol. 35, No. 7, 3600–3624. doi:10.1177/0958305X231171702.
22. Zhang, H. (2023). Global Energy Demand and Forecasting: A Managerial Economics Perspective, *Advances in Economics, Management and Political Sciences*, Vol. 19, No. 1, 144–149. doi:10.54254/2754-1169/19/20230130.
23. ADB. (2022). MT-GT Strategic Framework and Roadmap 2022–2026, Asian Development Bank, 65–96.
24. ADB. (2021). *Implementation Blueprint 2017-2021*, Asian Development Bank.

25. Dong, Y., Li, X., Wu, J., and Liu, Y. (2020). International Competitiveness and Potential for Trade Cooperation on Clean Energy between the USA and China, *E3S Web of Conferences*, Vol. 194, 02007. doi:10.1051/e3sconf/202019402007.

26. Ağbulut, Ü., Gürel, A. E., and Biçen, Y. (2021). Prediction of Daily Global Solar Radiation Using Different Machine Learning Algorithms: Evaluation and Comparison, *Renewable and Sustainable Energy Reviews*, Vol. 135, 110114. doi:10.1016/j.rser.2020.110114.

27. Attanasi, E. D., and Coburn, T. C. (2023). Random Forest, 1182–1185. doi:10.1007/978-3-030-85040-1_265.

28. Noviandy, T. R., Idroes, G. M., Mohd Fauzi, F., and Idroes, R. (2024). Application of Ensemble Machine Learning Methods for QSAR Classification of Leukotriene A4 Hydrolase Inhibitors in Drug Discovery, *Malacca Pharmaceutics*, Vol. 2, No. 2, 68–78. doi:10.60084/mp.v2i2.217.

29. Garcia-Ramirez, I.-A., Calderon-Mora, A., Mendez-Vazquez, A., Ortega-Cisneros, S., and Reyes-Amezcua, I. (2022). A Novel Framework for Fast Feature Selection Based on Multi-Stage Correlation Measures, *Machine Learning and Knowledge Extraction*, Vol. 4, No. 1, 131–149. doi:10.3390/make4010007.

30. Yang, X., Zhang, C., Li, Q., and Cheng, J.-H. (2023). Physicochemical Properties of Plasma-Activated Water and Its Control Effects on the Quality of Strawberries, *Molecules*, Vol. 28, No. 6, 2677. doi:10.3390/molecules28062677.

31. Qi, S., Sun, Y., Yin, P., Zhang, H., and Wang, Z. (2021). Mobile Phone Use and Cognitive Impairment among Elderly Chinese: A National Cross-Sectional Survey Study, *International Journal of Environmental Research and Public Health*, Vol. 18, No. 11, 5695. doi:10.3390/ijerph18115695.

32. Sahriman, S., and Yulianti, A. S. (2023). Statistical Downscaling Model With Principal Component Regression and Latent Root Regression To Forecast Rainfall in Pangkep Regency, *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, Vol. 17, No. 1, 0401–0410. doi:10.30598/barekengvol17iss1pp0401-0410.

33. Ulhaq, M. Z., Farid, M., Aziza, Z. I., Nuzullah, T. M. F., Syakir, F., and Sasmita, N. R. (2024). Forecasting Upwelling Phenomena in Lake Laut Tawar: A Semi-Supervised Learning Approach, *Infolitika Journal of Data Science*, Vol. 2, No. 2, 53–61. doi:10.60084/ijds.v2i2.211.

34. Bintoro, P., Zulkifli, Z., Fitriana, F., and Sukarni, S. (2024). Deep Convolutional Neural Network for Lampung Character Recognition, *Bulletin of Electrical Engineering and Informatics*, Vol. 13, No. 4, 2471–2484. doi:10.11591/eei.v13i4.6734.

35. Hahs-Vaughn, D. L. (2021). Multivariate Statistics, *Applied Multivariate Statistical Concepts*. doi:10.4324/9781315816685-6.

36. Zhang, H., Jiang, B., Zhang, H., Li, P., Wu, M., Hao, J., and Hu, Y. (2024). Water–Rock Interaction Mechanisms and Hydrochemical Evolution in the Underground Reservoirs of Coal Mines, *ACS Omega*, Vol. 9, No. 43, 43834–43849. doi:10.1021/acsomega.4c06809.