Vol. 15 No. 1
January 2024

DOI : https://doi.org/10.18196/jsp/v15i1.360
http://journal.umy.ac.id/index.php/jsp

Towards Sustainable Energy Policies: Machine Learning Applications in Projecting Bio Solar Consumption in Indonesia

AFFILIATION:
Department of Industrial Engineering, Universitas Indonesia

CORRESPONDENCE:
atiq.mujtaba90@gmail.com ;
Komarudin74@ui.ac.id

HOW TO CITATE:
Mujtaba, A., & Komarudin. (2024). Towards Sustainable Energy Policies: Machine Learning Applications In Projecting Bio Solar Consumption In Indonesia. *Jurnal Studi Pemerintahan, 15*(1).104-131

ARTICLE HISTORY:
Received : November 30, 2023
Revised : January 27, 2024
Accepted : February 29, 2024

ATIQ MUJTABA, KOMARUDIN

ABSTRACT

This paper explores the application of machine learning (ML) techniques to project the future consumption of bio solar energy in indonesia, aiming to inform and guide policy decisions in the energy sector. The transition to renewable energy sources is crucial for sustainable development, especially in emerging economies like indonesia, which has shown a growing interest in bio solar energy. This research method uses Quantitative Research with Linear Regression and Sarima approaches. We employed several ML models, using Phyton which analyse with Multiple Linear Regression, Lasso Regression and Sarima, to analyze historical data on energy consumption, economic indicators, demographic changes, and technological advancements. Our findings indicate that ml models can effectively predict bio solar consumption trends, highlighting the influence of economic growth, urbanization, and technological innovation on renewable energy adoption. The models suggest an increasing trajectory in bio solar consumption, driven by policy incentives, technological advancements, and a growing awareness of environmental issues. The accuracy of ml predictions is contingent upon the availability and quality of data. Furthermore, the projections may not account for unforeseen economic or technological changes. Future research should focus on incorporating more dynamic data sources and exploring the impact of policy changes on renewable energy adoption. In conclusion, leveraging machine learning for policy projection offers a promising approach to support the growth of bio solar consumption in indonesia. This study provides a foundation for future research and highlights the potential of ml in crafting informed, effective energy policies.

Keywords: Machine Learning, Policy Projection, Bio Solar Consumption, Indonesia

ABSTRAK

Paper ini mengeksplorasi penerapan teknik machine learning (ML) untuk memproyeksikan konsumsi energi bio-solar di Indonesia di masa depan, yang bertujuan untuk memberikan informasi dan memandu pengambilan kebijakan di sektor energi. Transisi ke sumber energi terbarukan sangat penting bagi pembangunan berkelanjutan, terutama di negara-negara berkembang seperti Indonesia, yang telah menunjukkan peningkatan minat terhadap energi biosolar. Metode penelitian ini menggunakan Penelitian Kuantitatif dengan pendekatan Regresi Linier dan Sarima. Kami menggunakan beberapa model ML, menggunakan Phyton yang menganalisis dengan Multiple Linear Regression, Lasso Regression, dan Sarima, untuk menganalisis data historis mengenai konsumsi energi, indikator ekonomi, perubahan demografi, dan kemajuan teknologi. Temuan kami menunjukkan bahwa model ml dapat secara efektif memprediksi tren konsumsi biosolar, menyoroti pengaruh pertumbuhan ekonomi, urbanisasi, dan inovasi teknologi terhadap adopsi energi terbarukan. Model-model tersebut menunjukkan adanya peningkatan konsumsi biosolar, didorong oleh insentif kebijakan, kemajuan teknologi, dan meningkatnya kesadaran akan isu-isu lingkungan. Keakuratan prediksi ml bergantung pada ketersediaan dan kualitas data. Selain itu, proyeksi tersebut mungkin tidak memperhitungkan perubahan ekonomi atau teknologi yang tidak terduga. Penelitian di masa depan harus fokus pada penggabungan sumber data yang lebih dinamis dan mengeksplorasi dampak perubahan kebijakan terhadap penerapan energi terbarukan. Kesimpulannya, pemanfaatan pembelajaran mesin untuk proyeksi kebijakan menawarkan pendekatan yang menjanjikan untuk mendukung pertumbuhan konsumsi biosolar di Indonesia. Studi ini memberikan landasan untuk penelitian di masa depan dan menyoroti potensi ml dalam menyusun kebijakan energi yang terinformasi dan efektif.

Kata Kunci: Machine Learning, Proyeksi Kebijakan, Konsumsi Bio Solar, Indonesia

INTRODUCTION

Energy is a basic need for many activities, regardless of whether a country is developed or developing. The proportion of commercial energy consumption caused by industrial energy use in developing countries is estimated to be around 45-50% (Sun & Chen, 2023). Controlling energy consumption is of great concern because of the interdependence between future global conditions and current choices. Effective management of energy resources has become an important concern for energy planners and policy makers (Aklilu, 2020). There is an increasing need for a comprehensive understanding of commercial and renewable energy sources, covering aspects such as quality, availability, and environmental impact (Maaouane et al., 2021).

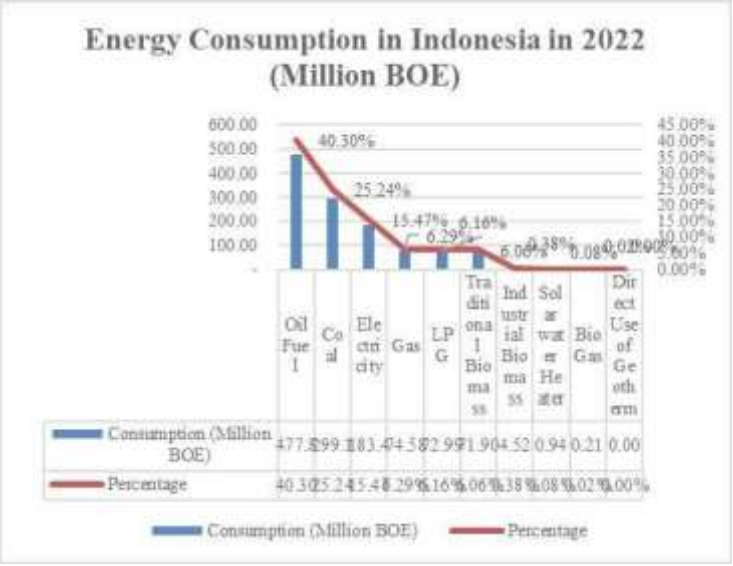

Figure 1. Energy Consumption in Indonesia in 2022 (Million BOE)
Source: (HEESI KESDM, 2022)

Energy availability is a key factor in supporting the economy and development of a country (Li et al., 2022). In Indonesia, energy consumption levels continue to increase in line with rapid economic growth and increasingly active economic activity. According to data from the Ministry of Energy and Mineral Resources (ESDM) in 2022, the level of energy consumption in Indonesia has increased from 754.404 million barrels of oil equivalent (BOE) in 2011 to 1,185.56 million BOE in 2022. The highest energy consumption in Indonesia is still dominated by fuel type energy which reached 40.30%. This energy consumption is used in various sectors, including industry, household, commercial and transportation (Rao et al., 2023). The Industrial sector is the largest contributor with 45.11% of total energy consumed, followed by the Transportation sector with 36.15%, the household sector with 13.62%, the commercial sector with 4.19%, and other sectors with 0.93 %. Graphs for energy consumption are seen in Figure 1 and Figure 2.

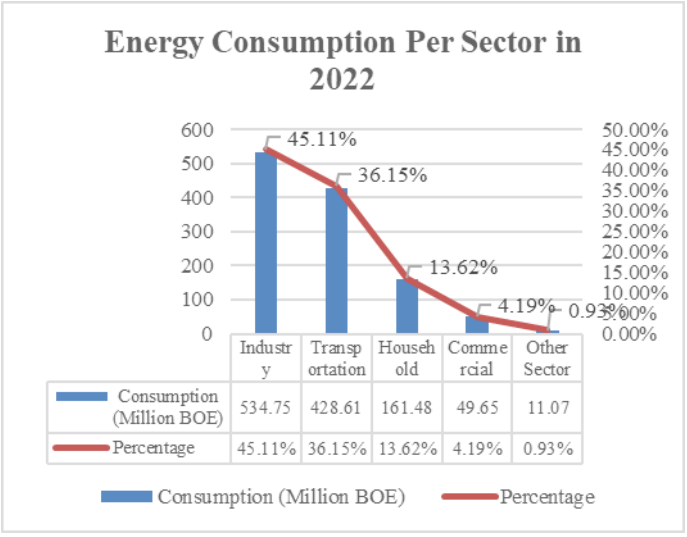

Figure 2. Energy Consumption Per Sector in Indonesia in 2022 (Million BOE)
Source: (HEESI KESDM, 2022)

Fuel oil is an important commodity in running a country's economy, fuel oil also has a big impact on the lives of many people (Zeng et al., 2023). In accordance with the mandate of Article 33 paragraph (2) and paragraph (3) of the 1945 Constitution of the Republic of Indonesia (UUD NKRI 1945), which states that production sectors which have an important role in the state and community life are supervised by the state and used to achieve maximum welfare for the people. In other words, in terms of energy supply, the Government must optimally ensure that public access to utilize energy sources is available as much as possible.

As for the types of fuel oil as referred to in Article 2 of Presidential Regulation Number 191 of 2014, BBM consists of Certain Types of Fuel (JBT), Special Types of Fuel for Assignments (JBKP) and General Types of Fuel (JBU). Based on Article 1 of Presidential Regulation Number 191 of 2014, JBT is fuel that is given a subsidy, JBKP is fuel that is distributed in the assigned area and is not given a subsidy, while JBU is general fuel that is

not given a subsidy. Based on Article 3 paragraph (1) Presidential Regulation Number 191 of 2014, certain types of fuel as referred to in Article 2 letter a consist of Diesel Oil (Gas Oil).

The government continues to adjust the mechanism for providing fuel subsidies (Liu & Wu, 2023; Sugiyono, 2008). One of the mechanisms used to regulate the amount of subsidized fuel is determining allocations/quotas. Quotas are determined to ensure fuel availability and maintain fuel price stability. Based on article 21 paragraph 5 of Presidential Decree 191/2014, the allocation (quota) for the volume of certain types of fuel is determined by the Regulatory Body. The Migas Downstream Regulatory Agency (BPH) is the institution responsible for determining the amount of Bio Solar Subsidy distributed throughout Indonesia. A comparison of quotas with the realization of subsidized fuel can be seen in Figure 3.

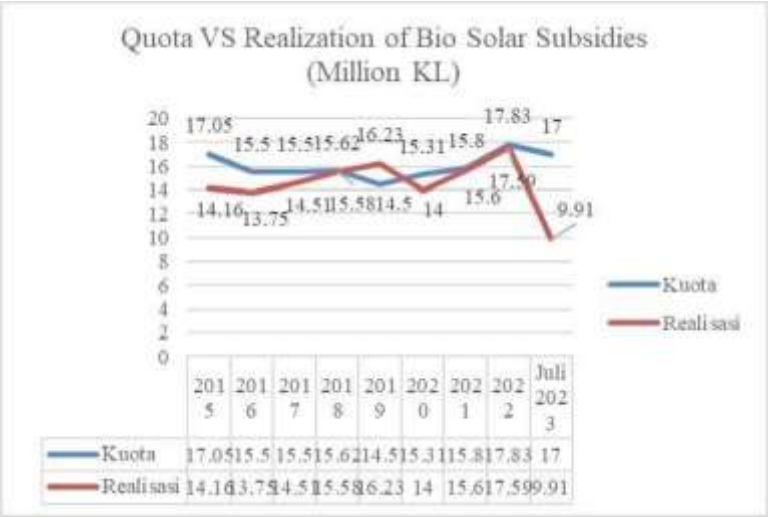

Figure3. vs Realization of Bio Solar Subsidies 2015-2023.
Source: (BPH Migas, 2023)

Therefore, in administering fuel to be able to achieve equitable fuel both in terms of availability, accessibility and affordability as mandated by the constitution and national energy policy, the Government is obliged to guarantee the avail-

ability and smooth distribution of fuel itself throughout Indonesia, which implementation is carried out by the Regulatory Body (Chanthawong et al., 2016a; Rahman et al., 2019).

Based on Figure 3, the distribution of Bio Solar Subsidy in 2019 experienced an over quota. This fact indicates that there is a need for a more actual projection model for Subsidized Bio Solar Fuel so that it will minimize the occurrence of over quotas for Subsidized Solar Fuel in the future. Data for 2022 shows that the amount of energy subsidies issued by the Government reaches 502.4 trillion rupiah, which is allocated for subsidies and compensation for fuel (53%), LPG gas (27%) and electricity (20%) (KESDM, 2022). Based on these data, most of the energy subsidy budget is allocated for fuel subsidies and compensation. Accuracy in projecting demand for subsidized Bio Solar fuel will influence the precise allocation of subsidized fuel quotas which has the potential to increase the efficiency of using the APBN for fuel subsidy spending.

Several literatures discuss fuel oil demand projections with quite varied results. The accuracy of estimates can be influenced by various factors such as economic growth, population growth, technological changes and energy prices by using the multiple linear regression model method and obtaining independent variables including population, price index, parity (Sahraei et al., 2021). purchasing power, gross domestic product, and household final consumption (Kristyadi et al., 2022).

Indonesia, a country with an archipelagic nature and a developing economy, has the task of dealing with increasing energy needs and managing environmental problems. Indonesia, as a member of the Association of Southeast Asian Nations (ASEAN), has implemented obligations to reduce greenhouse gas emissions and advance sustainable energy alternatives (Setiyawan et al., 2022). The bio-diesel fuel market in Indonesia faces various problems, such as price instability, policy inconsistencies, and lack of infrastructure, even though there is potential

profit (Handra & Hafni, 2017). Subsidies are often used as a policy instrument to stimulate consumption and increase the competitiveness of bio-diesel fuel compared to fossil fuels. However, providing these subsidies can put significant pressure on a country's fiscal resources and can have unexpected economic and ecological impacts (Kristyadi et al., 2022).

The increasing need for energy in Indonesia is putting pressure on the supply of conventional fossil fuels, requiring a transition to more sustainable energy sources. In contrast to conventional fossil fuels, bio-diesel fuel has the advantageous characteristics of being biodegradable and reducing greenhouse gas emissions, this makes it a very attractive alternative to encourage sustainable development (Rao & Dimitropoulos, 2023). The Indonesian government always provides fossil fuel subsidies to keep prices affordable for consumers. However, this methodology has come under intense scrutiny due to its detrimental ecological impacts and significant pressure on government fiscal resources. To encourage the transition towards biodiesel sources, the government has implemented subsidy initiatives that are clearly targeted at biodiesel fuel (Chanthawong et al., 2016b). Various policy instruments have been used or suggested to scale up the biodiesel fuel industry, including direct cash incentives, tax exemptions, and mandatory production targets. The extent to which these policies contribute to increased sustainable use of biodiesel fuel is still not fully understood. A few scientific investigations have been carried out to explore the impact of subsidies on the use of bio-diesel sources. Therefore, it is very important to estimate the consumption patterns of subsidized bio diesel fuel in Indonesia to provide valuable insights in making sustainable policy decisions. This study will examine data on subsidized bio-diesel fuel consumption in Indonesia over the past decade from 2014 to 2022. This analysis will focus on the next decade to one and a half decades, covering factors such as economic expansion, technological developments, and energy costs all over the world (Kim, 2013). The aim of this research is to

provide a reference for policy makers, business stakeholders and academics who are interested in studying the demand for Bio Solar Subsidies in Indonesia. Research on bio diesel consumption policies in Indonesia usually aims to evaluate, understand, and improve renewable energy policies in the country, with a special focus on bio diesel. Bio diesel is diesel fuel mixed with biofuel, which is expected to reduce dependence on fossil fuels, reduce greenhouse gas emissions, and support local economic development using domestic resources. The purpose of this introduction is to provide the reader with a comprehensive understanding of the importance of the research, its goals, and the methodology used to achieve those goals.

METHODOLOGY

This research method uses Quantitative Research with Linear Regression and Sarima approaches. The research method using linear regression is a statistical approach used to study the relationship between one or more independent variables (predictors) and dependent variables (criteria). This method is often used in various research fields such as economics, social sciences, biology, and others to predict the value of the dependent variable based on known information from the independent variable. Linear regression is a statistical method used to analyze the linear relationship between two variables. This method can be used to predict the value of one variable (dependent variable) based on the value of another variable (independent variable). Simple linear regression involves two variables, while multiple linear regression involves more than two independent variables. The basics of simple linear regression theory are as follows:

Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

a. Y is the dependent variable that you want to predict.
b. X is the independent variable.
c. β_0 is the intercept (cut point) of the regression line.

d. β_1 is a regression coefficient that shows how much Y changes due to a one unit change in X.
e. ε is random error.

The main goal is to find the best regression line that minimizes the prediction error (å). This can be measured by methods such as the least squares method. Minimize the sum of squared prediction errors (SSE) to find the optimal values of â€ and â .

\[SSE = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \]

Coefficient Estimates:

- \[\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i= 1}^{n} (x_i - \bar{x})^2} \]
- \[\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \]

\[\bar{y} \] is the average of the Y values, and \[\bar{x} \] is the average of the X values.

- Coefficient of Determination (\(R^2\)): Measures how well the model can explain the variability of the data.

\[R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \]

Linear Regression Assumptions:

a. The assumption about the error distribution (å) is normal.
b. Homoscedasticity: Constant error variance for all X values.
c. No multicollinearity: There is no strong relationship between the independent variables.
d. No autocorrelation: There is no pattern in the error distribution.

Test the significance of the coefficient to determine whether the independent variable has a significant influence.

Linear regression is a tool commonly used in statistical and econometric analysis to understand the relationship between variables. This model provides the basis for a variety of more complex regression techniques and is used in a variety of scien-

tific disciplines such as economics, social sciences, and natural sciences.

This study presents an integrated approach using six Machine Learning (ML) algorithms and mathematical programming to predict energy demand in Iran until 2040. Data from electricity generation, fuel consumption, and seven major energy consuming sectors were collected. The ML algorithm and prediction accuracy index was evaluated in each sector (Li et al., 2022). An optimization model for improving prediction accuracy is introduced, which is executed by two algorithms PSO and Grey-Wolf Optimizer for different sectors. The results show that the proposed method has higher prediction accuracy than other ML algorithms, with the PSO algorithm predicting an increase in total energy requirements of 75.65% compared to 2018, and the Grey-Wolf Optimizer algorithm predicting an increase of 82.94%. The study shows that the integrated machine learning algorithm can predict with high prediction accuracy, and the integrated model can be used for future research with different prediction algorithms.

Even though there is a lot of research on energy/fuel demand projections, there is still a lack of research that explicitly examines the projected demand for bio-diesel fuel in the context of subsidies in Indonesia. Considering the different socio-economic and environmental conditions in a country, especially Indonesia, the initial hypothesis is that the independent variable, the comparison of subsidized and non-subsidized prices, will have a significant influence on the demand for subsidized Bio Solar fuel in Indonesia, including the independent variable GDP per sector that is the consumer user. fuel Bio Solar subsidies., This statement underscores the importance of the subject matter and places current research in a broader academic and policy context.

This study was conducted to measure the influence of various economic and social indicators on the realization of the Bio Solar Subsidized BBM. This research will compare the 3 meth-

ods most often used in energy/fuel demand projections, namely regression models, SARIMA models (time series) and artificial neural network modeling. Economic factors were selected based on a literature review of previous research and studies. The projection model aims to provide information in determining future subsidized fuel quotas based on projected demand for Bio Solar Subsidies. It is hoped that the validated model can be used as a consideration in the process of determining subsidized fuel quotas in the future.

Lasso (Least Absolute Shrinkage and Selection Operator) is a regression method used for variable selection and model regularization in statistics and machine learning. Lasso was developed as an extension of linear regression to overcome multicollinearity problems and to produce a more parsimonious model. The objective function for Lasso is as follows:

\[\text{minimize } \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}) ^2 + \lambda \sum_{j=1}^{p} |\beta_j| \]

a. \(n\) is the number of observations.
b. \(p\) is the number of independent variables.
c. \(y_i\) is the value of the dependent variable for the \(i\)th observation.
d. \(x_{ij}\) is the value of the \(j\)th independent variable for the \(i\)th observation.
e. \(\beta_0, \beta_1, \ldots, \beta_p\) are the regression coefficients to be estimated.
f. \(\lambda\) is a regularization parameter that controls how strong the regularization is applied. The larger \(\lambda\), the greater the effects of variable shrinkage and selection.

Regularization L1:

The importance of Lasso lies in the term \(\lambda \sum_{j=1}^{p} |\beta_j| \), which is part of L1 regularization. This term applies “shrinkage” to the regression coefficients and effectively pushes some of the coefficients to zero. Thus, Lasso

not only produces predictions, but also performs feature selection. Lasso can automatically remove variables that have a small influence on the dependent variable, making it easier to interpret the model and increasing model stability when the number of features (independent variables) is very large. In Lasso, hypothesis testing is related to zero coefficients. If the coefficient is considered zero, the variable is considered not to make a significant contribution to the model. Parameter selection \(\lambda\) in Lasso is often done through cross-validation techniques, where the model is tested on different subsets of data to evaluate the model's performance on data not used during training.

Lasso has unique properties in selecting variables, whereas Ridge regression is more likely to retain all variables in the model. Lasso is a useful tool in situations where there are many independent variables and some of them may be irrelevant or correlated with each other. By introducing L1 regularization, Lasso helps build models that are more parsimonious and easier to interpret.

This study uses primary and secondary data. Dependent Variable Data, namely Sales of Subsidized Bio Solar, is data owned by BPH Migas and sourced from the SILVIA BPH Migas website and the Ministry of Energy and Mineral Resources. Independent variable data is also used according to study needs originating from various government agencies, including the Central Statistics Agency (BPS), Pertamina, and BI. This research focuses on the relationship between the realization of Bio Solar Subsidy Fuel and independent variables that influence Household Expenditures; Government Expenditure; PKP (Agriculture, Forestry, Fisheries; TP (transportation and warehousing); JKS (Health Services and social activities; KRT (Household consumption); CPI (Consumer Price Index); Number of Passenger Cars; Number of Buses; Number of Trucks; GDP; and Price of diesel vs Pertamina dex.

RESULT AND DISCUSSION

The highest energy consumption in Indonesia is dominated by oil fuel at 40.30%. Followed by coal at 25.24%. Energy consumption is used in various sectors, including industry, household, commercial and transportation. The industrial sector is the largest contributor, namely 45.11% of total energy consumption. Followed by transportation at 36.15%, bio-diesel distribution in 2020 decreased due to Covid-19. In 2021 and 2022, realization vs quota will almost reach 100%, while in 2019 there will be an excess quota. The impact of over quota is that the assignment business entity is not paid the subsidy. In 2022, most energy subsidies will be allocated to fuel subsidies and compensation. subsidies and compensation for fuel (53%), LPG gas (27%) and electricity (20%). It is hoped that accurate projections can increase the efficiency of spending on fuel subsidies, and minimize losses for assigned Business Entities from not paying subsidies due to over quota. Based on the table of statistical values from the Kolmogorov-Smirnov test, the statistical results are compared with the table values of KS N=35, á =
0.05 is 0.224, and P-Value > 0.05 is normally distributed. So that all independent variable data is normally distributed (table 1)

This test compares the cumulative distribution of the data sample with the theoretical cumulative distribution expected in the case of a normal distribution. In other words, the K-S test measures how well the sample data distribution matches the expected normal distribution. This indicates that the sample size used in the research or analysis is 35. The K-S test is sensitive to sample size; therefore, the sample size needs to be known for correct interpretation of the results. á = 0.05: This is the significance level chosen for statistical tests. The significance level á is the probability of rejecting the null hypothesis (in this case, that the data follows a normal distribution) when in fact the null hypothesis is true. A significance level of 0.05 or 5% is a commonly used standard, meaning that the risk of committing a type

Table 1. Normality test for independent variables

	Column	Statistic	P-Value
0	Subsidized Solar Sales	0.109336	0.855640
1	Household Expenditures	0.098567	0.923770
2	Government Spending	0.104200	0.890712
3	PKP (Agriculture, Forestry, Fisheries)	0.075616	0.993376
4	TP (Transportation and Warehousing)	0.115562	0.807878
5	JKS (Health Services and Social Activities	0.143845	0.559884
6	KRT (Household Consumption)	0.109651	0.853354
7	CPI (Consumer Price Index for Inflation)	0.124218	0.734731
8	Number of Passenger Cars	0.128787	0.694284
9	Number of Buses	0.133366	0.653220
10	Number of Trucks	0.145462	0.545791
11	GDP	0.095016	0.941395
12	Solar VS Pertadex	0.213715	0.133380

I error (rejecting the null hypothesis when it should be accepted) is 5%. In the context of the K-S test, the critical value is the threshold value determined by the sample size and significance level. For N=35 and á=0.05, the critical value is 0.224. This means that if the K-S test statistic (which measures the maximum difference between a sample's cumulative distribution and a normal distribution) is smaller than 0.224, then the difference is not statistically significant at the 5% significance level. P-Value > 0.05: P-value is the probability of getting an observation result that is extreme or more extreme than what is obtained, if the null hypothesis is true. In this context, a P-Value greater than 0.05 indicates that there is not enough statistical evidence to reject the null hypothesis. In other words, there is not enough evidence to say that the data is not normally distributed. Based on the Kolmogorov-Smirnov test with a sample size of 35 and a significance level of 0.05, if the test statistical value is smaller than the critical value of 0.224 and the P-Value is greater than 0.05, then we can conclude that there is not sufficient evidence to reject the hypothesis that the data is normally distributed. This means that the data for all independent variables are considered to have a normal distribution. This is important in statistical analysis

because many parametric statistical techniques require the assumption that the data is normally distributed.

Multicollinearity test for independent and dependent variables with a VIF. dropping variables for independent variables to obtain variables that will be used in the linear regression model as follows: Government Expenditure, PKP (Agriculture, Forestry, Fisheries), JKS (Health Services and social activities, Consumer Price Index, Subsidized Diesel Sales, and Diesel Price vs. Pertamina Dex.

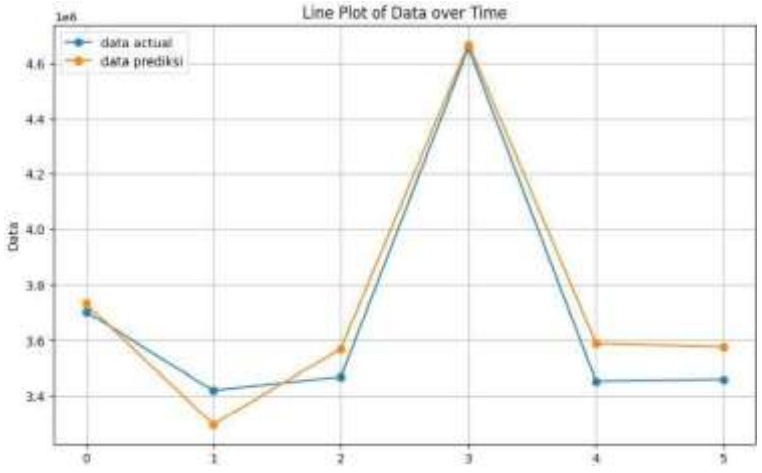

Figure 4. Result of Forecasting with Regression Linear

R-squared (R2) Score: 0.9498. R-squared is a statistical measure in a regression model that indicates the proportion of variance for a dependent variable that can be explained by the independent variables in the model. The R-squared value is between 0 and 1, with higher values indicating a better fit between the model and the data. The R-squared score of 0.9498 (or about 95%) is very high, indicating that the linear regression model has a very good fit. This means that almost 95% of the variance of the dependent variable can be explained by the model. MAPE is a metric for assessing the accuracy of model predictions. MAPE measures the average absolute error in percentage, which allows

easy interpretation in terms of relative error. The MAPE score of 2.4844% indicates that the model, on average, has an error of about 2.48% from the actual value. In many contexts, a MAPE below 5% is considered accurate, so this score indicates good performance of the regression model in terms of prediction accuracy. In conclusion, the analyzed linear regression model appears to provide an excellent fit and high prediction accuracy based on the reported R-squared and MAPE values.

Table 2. Multiple Linear Regression Model

Variable	VIF
Government Spending	1.818827
PKP (Agriculture, Forestry, Fisheries),	2.073042
JKS (Health Services and social activities,	2.911910
CPI (Consumer Price Index),	1.229131
Price of diesel vs pertadex	1.843292

The table 2 shown that the Variable Inflation Factor (VIF) values for several economic and social variables in a particular country or study. Here is the explanation:

1. Government Spending: The VIF figure of 1.81827 shows the level of correlation between government spending and other variables in the model. This value is below 5, which is generally considered the threshold at which multicollinearity (high correlation between predictor variables) becomes a serious problem.
2. PKP (Agriculture, Forestry, Fisheries): With a VIF of 2.073042, this indicates a moderate correlation between the agricultural, forestry and fisheries sectors with other variables in the model, but is still considered acceptable.
3. JKS (Health services and social activities): This has the highest VIF in the table, 2.911910, indicating that there is a higher correlation between health services and social activities with

other variables compared to other variables in the table. However, this value is still below 5.

4. CPI (Consumer Price Index): VIF of 1.229131 indicates that the consumer price index has a relatively low correlation with other variables in the model.
5. Price of diesel vs pertadex: The VIF value of 1.843292 shows a relatively low correlation between the price comparison of diesel and pertadex with other variables in the model.

In general, these VIF values indicate that there is no serious multicollinearity problem between these variables. This is important in linear regression because multicollinearity can obscure the interpretation of predictor variable coefficients and reduce the accuracy of the model. It should be noted that this interpretation is valid if the regression model used is appropriate and the regression assumptions have been met. Additionally, these values must be interpreted in the context of the data and scales used. Large coefficients can occur because the scale of the independent or dependent variable is large or because the relationship is very strong. Very large or very small coefficients can indicate problems with the data or model, such as multicollinearity or the influence of outliers.

Table 3. Lasso Regression Model

Independent Variable	Coeficient Value
TP (transportation and warehousing)	9.31708066e+00
JKS (Health services and social activities)	2.70210710e+01
KRT (Household Consumption)	4.81691540e+00
IHK (Consumer Price Index)	-3.59073281e+03
Number of Buses	-4.80093130e+01
Price of diesel vs pertadex	-2.73089939e+06

Lasso regression can help select the most important independent variables for forecasting. A regression coefficient that is close to zero indicates that the related independent variable does not have a significant influence on the dependent variable, then the variables are dropped to produce the independent variables TP, JKS, KRT, CPI, Number of Buses, and price of diesel vs Pertadex according to the table 3. The provided coefficients are from a statistical model showing how different variables affect a certain outcome. Here is a brief explanation of each variable and its coefficient:

1. TP (Transportation and Warehousing): A coefficient of approximately 9.32 suggests that an increase in transportation and warehousing activities is associated with a significant positive effect on the outcome variable.
2. JKS (Health Services and Social Activities): With a coefficient of approximately 27.02, this indicates a very strong positive relationship, suggesting that improvements or increases in health services and social activities have a substantial positive impact.
3. KRT (Household Consumption): A coefficient of about 4.82 indicates a positive relationship, meaning household consumption positively affects the outcome, though its impact is less than TP and JKS.
4. IHK (Consumer Price Index): The coefficient of approximately -3590.73 suggests a strong negative effect on the outcome variable, indicating that as the consumer price index increases, the outcome variable significantly decreases.
5. Number of Buses: With a coefficient of about -48.01, it suggests that an increase in the number of buses has a negative impact on the outcome variable, though the impact is smaller compared to the consumer price index.
6. Price of Diesel vs. Pertadex: The coefficient of approximately -2,730,899.39 indicates an extremely strong negative impact on the outcome, suggesting that increases in the price differential between diesel and pertadex have a highly significant negative effect.

These coefficients provide insights into how each independent variable influences the dependent variable, with positive coefficients indicating a positive impact and negative coefficients indicating a negative impact.

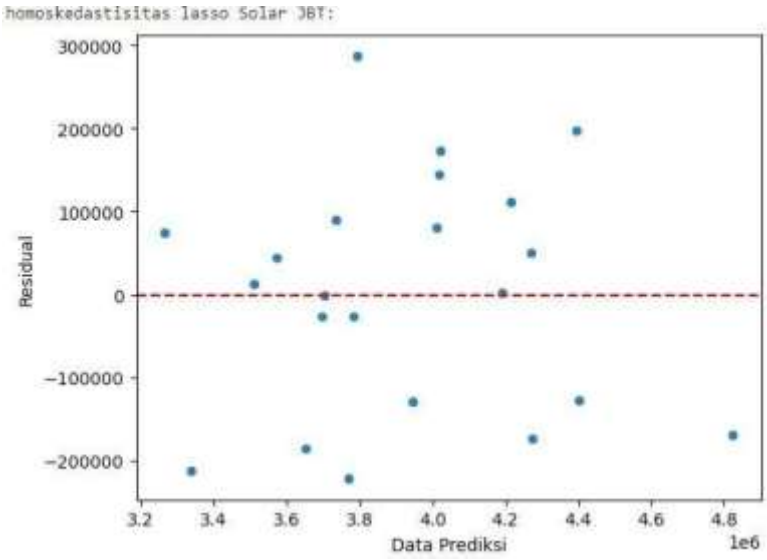

Figure 5. Result of Homoskedastisitas

In Figure 5, the Lasso regression model for data related to Bio Solar Subsidy. This graph is used to check homoscedasticity, namely the assumption that the variance of the regression model error is constant at all levels of predicted values. It displays the values predicted by the model. These numbers range from about 3.2 million to almost 5 million, which can represent predictions from the model, such as estimates of diesel consumption. It displays the residual, which is the difference between the observed value and the value predicted by the model. A positive residual means the model has predicted a lower value than the actual (underestimate), while a negative residual means the model has predicted a higher value than the actual (overestimate). Homoscedasticity is indicated by a uniform distribution of re-

siduals along the level of the predicted value. In this graph, there seems to be some increase in residual variability as a function of predicted value as the points appear to be spread wider at higher predicted values. However, without a clear or systematic pattern such as a funnel or certain shape, it cannot be concluded with certainty that there is non-homoscedasticity. This indicates a residual value of zero. The residuals will ideally be randomly distributed around this line, with no clear upward or downward pattern.

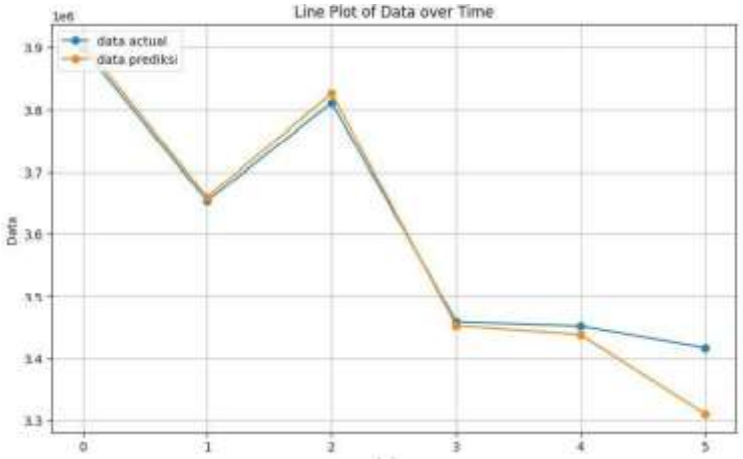

Figure 6. Forecasting Result using Lasso Regression

The R Squared value in Figure 6 shows the number 0.9434. The R-squared score is a measure of how well the model predictions match the actual data. R2 values range from 0 to 1, with values close to 1 indicating that the model has a very good fit to the data. With an R2 score of 0.9534, this indicates that your Lasso regression model explains approximately 95.34% of the variance of the dependent variable. This is an excellent fit, indicating that the model can predict with high accuracy based on its independent variables. In figure 6 the Mean Absolute Percentage Error (MAPE) shows 0.7429%. MAPE is a metric that evaluates a model's prediction error as a percentage. A low MAPE

indicates a smaller error rate between the value predicted by the model and the actual value. A MAPE score of 0.7429% indicates that on average, your Lasso model makes less than 1% error in the predictions it makes, indicating very high accuracy.

These two metrics together indicate that the Lasso regression model you are using is performing very well in the sample of data tested. A high R2 indicates a strong fit between the model and the observed data, and a low MAPE indicates a high degree of accuracy in the model predictions.

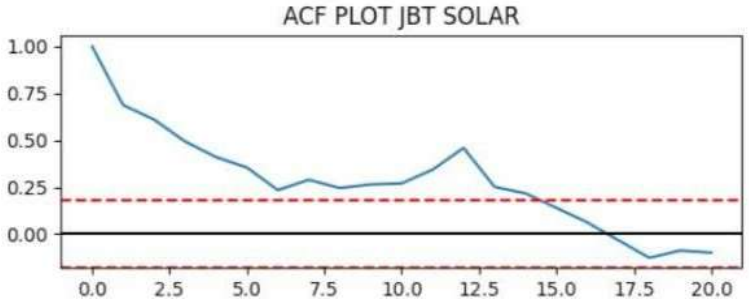

Figure 7. Autocorrelation Function Result of Phyton Analysis Forecasting Bio Solar Subsidy

The plot you uploaded is an Autocorrelation Function (ACF) plot for Bio Solar Subsidy data. ACF is used to measure how strongly a time series is correlated with itself at different lags. Figure 7 shows correlation values that range from -1 to 1. Values close to 1 indicate a strong positive correlation, while values close to -1 indicate a strong negative correlation. A value of 0 indicates no correlation. Lag indicates the number of time periods that are measured as 'lag'. In the context of a time series, this 'lag' can mean minutes, hours, days, etc., depending on the data. Dashed Red Line This marks the significance limit for the correlation. If the ACF line exceeds this red line, the correlation at that lag is considered statistically significant. In this graph, it appears that the ACF value exceeds the significance limit in the first few lags, which indicates a significant correlation. The rapidly decreasing

pattern in the initial lags indicates that the correlation between observations decreases rapidly as the lag increases. However, there is a spike at a point further away, indicating a higher correlation at that lag. From this ACF figure 7, we can draw conclusions about the memory properties of Bio Solar Subsidy data. The memory referred to here is the extent to which past values influence future values. This is important, for example, in time series modeling for forecasting purposes, where we want to know how much 'lag' we have to account for to make accurate predictions.

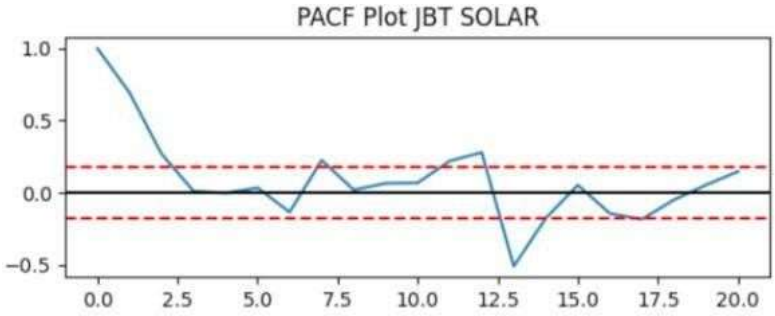

Figure 8. Partial Autocorrelation Function Result of Phyton Analysis Forecasting Bio Solar Subsidy

The plot you uploaded is a Partial Autocorrelation Function (PACF) plot for Bio Solar Subsidy data. PACF measures the partial correlation between a time series and itself at a given lag, controlling for correlation at all shorter lags. In other words, PACF at lag k is the correlation between the series at time t and time t-k, after removing the correlation from all shorter lags. This differs from ACF, which measures the total correlation between time series at different lags. Partial Correlation) shows the partial correlation value, which can range from -1 to 1. Values close to 1 or -1 indicate a strong, positive, or negative partial correlation, while values close to 0 indicate no partial correlation. The Lag axis shows the amount of lag, the same as in the ACF plot. the blue lines show the PACF values at different lags. A significant PACF value (i.e., outside the dashed red boundary line) indicates that there is a statistically significant correlation between observations at that lag interval, after controlling for correlation

at shorter lags. Dashed Red Line The dotted red horizontal line indicates the standard error limit for the partial correlation. If the PACF line is outside this limit, it is considered that there is a significant partial correlation at that lag.

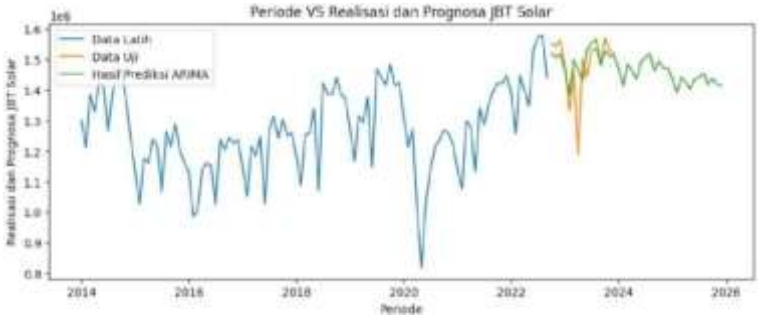

Figure 9. SARIMA Model Regression Period VS Consumption and Forecasting Bio Solar Subsidy

Figure 9 displays a comparison between actual and predicted data from the SARIMA model for Bio Solar Subsidy over a certain time. Displays time from 2014 to 2026, which may represent years or may be shorter time intervals, depending on the data frequency used. Indicates the measured value, possible consumption, or production of Diesel in unspecified units, but appears on a scale of 0.8 million to 1.6 million. It shows the historical data used to 'train' or calibrate the ARIMA model. From figure 9, it seems that the SARIMA prediction line quite follows the trend of the test data, which indicates that the model has good prediction capabilities. SARIMA predictions continue to follow historical patterns and do not adapt to sudden changes that may occur in test data; this is a general characteristic of SARIMA models that use historical information to make predictions. The forecast period shown by the green line extends beyond the most recent data available, providing an estimate of how Solar consumption or production may change in the future. Fluctuations in the test data (orange line) and model predictions (green line) show the variability in the actual data and how the SARIMA model attempts to adapt to that variability. The use of SARIMA models in this context may be to under-

stand Solar consumption or production patterns and try to predict those trends for future planning and strategic decision making. The model's performance in producing accurate predictions is critical to ensuring that the planning is reliable.

Table 4. MAPE Model

Bio Solar Subsidi				
Regresi Linier		Lasso		SARIMA
R Square	MAPE	R Square	MAPE	MAPE
0,9498	2,484%	0,9434	0,743%	2,251%

Source: phyton forecast (2024)

Table 4 which shows part of a statistical analysis table for the variable "Bio Solar Subsidies". In this table, there appear to be three analysis methods represented, namely "Linear Regression", "Lasso", and "SARIMA". However, only the "Lasso" method has the value displayed:

1. R Square: The R Square value of 0.9434 indicates that the Lasso model can explain around 94.34% of the variability of the dependent variable data, which in this case seems to be Bio Solar Subsidies. This is a very high value, indicating a good fit between the model and the data.
2. MAPE (Mean Absolute Percentage Error): MAPE of 0.743% is a measure of the average error in predictions, expressed as a percentage. This very low MAPE value shows that the predictions made by the Lasso model are very accurate, with an average prediction error of less than 1%.
3. The Linear Regression and SARIMA methods do not appear to show the values in table 4. Typically, Linear Regression will show how well the dependent data (in this case, Bio Solar subsidies) can be explained by the independent variables in the linear model, and SARIMA (Seasonal Autoregressive Integrated Moving Average) is a method used to analyze and predict time series data that has seasonal patterns.

Table 5. Prediction result

Year	Quartal	Prediction Multiple Linear Regression	Prediction Lasso Regression	Prediction SARIMA
2024	Q1	3.988.420	3.653.591	4.369.458
2024	Q2	4.251.141	3.773.414	4.391.658
2024	Q3	4.319.649	4.088.557	4.484.065
2024	Q4	4.670.931	4.380.119	4.436.645

Source: Phyton Forecasting (2024)

Table 5 shows all the predictions are for the year of 2024. Table 5 shows the predicted values using the Multiple Linear Regression model, which is a basic form of predictive analysis that assumes a linear relationship between the input variables and the single output variable. The prediction lasso regression is like linear regression but with a regularization factor that penalizes the absolute size of the coefficients, often resulting in some coefficients being shrunk to zero, effectively selecting a simpler model that may avoid overfitting. Also, SARIMA prediction stands for Seasonal Autoregressive Integrated Moving Average, which is particularly useful for forecasting data with seasonal patterns and non-stationary characteristics were data show trends or heteroscedasticity. In Q1 of 2024, the Multiple Linear Regression model predicts a value of 3,988,420, the Lasso Regression model predicts a value of 3,653,591, and the SARIMA model predicts a value of 4,369,458. In Q2 of 2024, the predictions are 4,251,141 from Multiple Linear Regression, 3,773,414 from Lasso Regression, and 4,391,658 from SARIMA. For Q3 of 2024, the predicted values are 4,319,649 from Multiple Linear Regression, 4,088,557 from Lasso Regression, and 4,484,065 from SARIMA. Finally, in Q4 of 2024, the predictions are 4,670,931 from Multiple Linear Regression, 4,380,119 from Lasso Regression, and 4,436,645 from SARIMA.

The predictions show a general increasing trend across the quarters for all three models. This could imply an expected growth in the variable being forecasted. The Multiple Linear Regression

model tends to give the highest predictions for each quarter except for Q1, where the SARIMA model is the highest. This might suggest a linear trend in the variable being forecasted or that this model is picking up on certain factors more strongly than the other models.

The Lasso Regression model generally gives the lowest predictions. This could indicate that it is the most conservative model, possibly due to the regularization effect that penalizes less important features. The SARIMA model, which is designed for time series data with seasonality, shows intermediate values compared to the other two models. This model's predictions could be considering potential seasonal patterns within the data. The differences in predictions might influence how resources are allocated or planning is conducted over the year. For instance, if the forecasts pertain to demand for a product, the government may plan to increase the quota of bio-diesel in anticipation of higher demand in later quarters.

CONCLUSION

This study has demonstrated the effectiveness of machine learning (ML) models in projecting the consumption patterns of bio solar in Indonesia. Our analysis, based on a comprehensive dataset spanning several years, indicates that ML algorithms can accurately predict future consumption trends based on a range of variables, including economic indicators, population growth, and technological advancements in renewable energy. The findings have significant implications for Indonesia's energy policy. By leveraging the predictive power of ML, policymakers can make informed decisions regarding the infrastructure investments needed to support bio solar consumption. This can lead to more sustainable energy policies that balance economic growth with environmental preservation. Additionally, our research suggests the need for policies that encourage the development and adoption of renewable energy technologies. The government could incentivize research in bio solar technologies and renewable en-

ergy, thereby enhancing Indonesia's energy security and reducing dependence on fossil fuels. Beyond policy implications, this study offers practical applications for energy providers and stakeholders in the renewable energy sector. By adopting ML models, these entities can better forecast demand, optimize supply chains, and develop targeted marketing strategies to increase bio solar adoption among consumers. Despite the promising results, this study has limitations. The accuracy of ML predictions is highly dependent on the quality and comprehensiveness of the data used. Furthermore, the models might not fully account for sudden policy changes or unforeseen global events affecting energy markets. Another limitation is the study's focus on Indonesia, which might limit the generalizability of the findings to other countries with different socio-economic and environmental contexts. Future research should aim to address the limitations mentioned above by incorporating more diverse and extensive datasets, including data from other countries to assess the model's applicability in different contexts. There is also a need to explore more advanced ML algorithms and techniques that could improve prediction accuracy and reliability. Investigating the impact of policy changes and global events on model predictions could provide more insights into how to make the models more resilient and adaptable. The usage of machine learning for policy projection of bio solar consumption in Indonesia offers a promising avenue for enhancing the sustainability and efficiency of the country's energy sector. As technology advances and more data become available, the potential of ML to inform and guide energy policy will only increase, leading to more informed decisions that ensure a sustainable and energy-secure future for Indonesia and beyond.

REFERENCES

Aklilu, A. Z. (2020). Gasoline and diesel demand in the EU: Implications for the 2030 emission goal. *Renewable and Sustainable Energy Reviews*, *118*. https://doi.org/10.1016/j.rser.2019.109530

Chanthawong, A., Dhakal, S., & Jongwanich, J. (2016a). Supply and demand of biofuels in the fuel market of Thailand: Two stage least square and three least square ap-

proaches. *Energy*, *114*, 431–443. https://doi.org/10.1016/j.energy.2016.08.006

Chanthawong, A., Dhakal, S., & Jongwanich, J. (2016b). Supply and demand of biofuels in the fuel market of Thailand: Two stage least square and three least square approaches. *Energy*, *114*, 431–443. https://doi.org/10.1016/j.energy.2016.08.006

Handra, N., & Hafni, H. (2017). Effect of Binder on Combustion Quality on EFB Bio-briquettes. *IOP Conference Series: Earth and Environmental Science*, *97*(1). https://doi.org/10.1088/1755-1315/97/1/012031

Kim, Y. (2013). Some recent empirical developments regarding the bio-industry: Using the Granger-causality test with Indonesia's palm oil data. *International Journal of Bio-Science and Bio- Technology*, *5*(5), 13–24 . https://doi.org/10.14257/ijbsbt.2013.5.5.02

Kristyadi, T., Permana, D. I., Sirodz, M. P. N., Saefudin, E., & Farkas, I. (2022). Performance and Emission of Diesel Engine Fuelled by Commercial Bio-Diesel Fuels in Indonesia. *Acta Technologica Agriculturae*, *25*(4), 221–228. https://doi.org/10.2478/ata-2022-0032

Li, Z., Zhou, B., & Hensher, D. A. (2022). Forecasting automobile gasoline demand in Australia using machine learning-based regression. *Energy*, *239*. https://doi.org/10.1016/j.energy.2021.122312

Liu, P., & Wu, J. (2023). Game Analysis on Energy Enterprises' Digital Transformation—Strategic Simulation for Guiding Role, Leading Role and Following Role. *Sustainability (Switzerland)*, *15*(13). https://doi.org/10.3390/su15139890

Maaouane, M., Zouggar, S., Krajaèiæ, G., & Zahboune, H. (2021). Modelling industry energy demand using multiple linear regression analysis based on consumed quantity of goods. *Energy*, *225*. https://doi.org/10.1016/j.energy.2021.120270

Rahman, S. A., Baral, H., Sharma, R., Samsudin, Y. B., Meyer, M., Lo, M., Artati, Y., Simamora, T. I., Andini, S., Leksono, B., Roshetko, J. M., Lee, S. M., & Sunderland, T. (2019). Integrating bioenergy and food production on degraded landscapes in Indonesia for improved socioeconomic and environmental outcomes. *Food and Energy Security*, *8*(3), 1–13. https://doi.org/10.1002/fes3.165

Rao, C., Zhang, Y., Wen, J., Xiao, X., & Goh, M. (2023). Energy demand forecasting in China: A support vector regression-compositional data second exponential smoothing model. *Energy*, *263*. https://doi.org/10.1016/j.energy.2022.125955

Rao, S., & Dimitropoulos, G. (2023). "Understanding the journey from A to Z": centering peer support perspectives to unveil the mechanisms and power of peer support. *Mental Health and Social Inclusion*. https://doi.org/10.1108/MHSI-02-2023-0016

Sahraei, M. A., Duman, H., Çodur, M. Y., & Eyduran, E. (2021). Prediction of transportation energy demand: Multivariate Adaptive Regression Splines. *Energy*, *224*. https://doi.org/10.1016/j.energy.2021.120090

Setiyawan, A., Novianto, A., Afkar, N. B. A., Chabib, F., Amelia, F. R., & Pratiwi, I. (2022). Diesel engine performance test using solar-dex and biodiesel (B30) on power and torque. *IOP Conference Series: Earth and Environmental Science*, *969*(1). https://doi.org/10.1088/1755-1315/969/1/012034

Sugiyono, A. (2008). *Pengembangan Bahan Bakar Nabati untuk Mengurangi Dampak Pemanasan Global Utilization of Biomass for Energy Sources View project Perencanaan energi nasional dan daerah View project*. https://www.researchgate.net/publication/275652084

Sun, J., & Chen, J. (2023). Digital Economy, Energy Structure Transformation, and Regional Carbon Dioxide Emissions. *Sustainability (Switzerland)*, *15*(11). https://doi.org/10.3390/su15118557

Zeng, Y., Xu, X., Zhao, Y., & Li, B. (2023). Impact of Digital Economy on the Upgrading of Energy Consumption Structure: Evidence from Mainland China. *Sustainability (Switzerland)*, *15*(7). https://doi.org/10.3390/su15075968