GIJES
Greenation International Journal of Engineering Science
+62 81210467572 https://research.e-greenation.org/GIJES greenation.info@gmail.com
GREENATION RESEARCH

DOI: https://doi.org/10.38035/gijes.v4i2

Evaluation of Linear Regression and Ridge Regression as Baselines for Predicting Indonesia's CO_2 Emissions

I Putu Santrisna[1], Dewa Gede Hendra Divayana[2], I Made Gede Sunarya[3].
[1]Universitas Pendidikan Ganesha, Bali, Indonesia, santrisna@student.undiksha.ac.id.
[2]Universitas Pendidikan Ganesha, Bali, Indonesia, hendra.divayana@undiksha.ac.id.
[3]Universitas Pendidikan Ganesha, Bali, Indonesia, sunarya@undiksha.ac.id.

Corresponding Author: santrisna@student.undiksha.ac.id[1]

Abstract: Indonesia's annual CO_2 emissions have increased alongside economic growth and rising energy demand, making short-term prediction relevant for environmental planning. This study evaluates Linear Regression and Ridge Regression as transparent baseline models for predicting Indonesia's total annual CO_2 emissions using public data from Our World in Data for 1990–2022. The predictors are GDP per capita, primary energy consumption, and population. The data were split chronologically, with 1990–2015 as training data and 2016–2022 as test data. Ridge Regression used standardized predictors, and its alpha parameter was selected on the training period using TimeSeriesSplit. Model performance was evaluated using MAE, RMSE, and R^2, and was compared with Naive Forecast and Linear Trend baselines. Linear Regression achieved the best performance on the test data with MAE of 19.181, RMSE of 21.384, and R^2 of 0.887. Ridge Regression produced MAE of 27.693, RMSE of 33.758, and R^2 of 0.719, while still outperforming the simple baselines. These results indicate that although regularization is theoretically motivated under multicollinearity, it did not improve, and in this case reduced, out-of-sample accuracy on a small annual dataset. The findings should be interpreted as predictive performance rather than causal relationships.

Keywords: *CO_2 Emissions, Linear Regression, Ridge Regression, Prediction.*

Indonesian Abstract: *Emisi CO_2 tahunan Indonesia meningkat seiring pertumbuhan ekonomi dan kenaikan permintaan energi, sehingga prediksi jangka pendek menjadi relevan untuk perencanaan lingkungan. Penelitian ini mengevaluasi Linear Regression dan Ridge Regression sebagai model baseline yang transparan untuk memprediksi total emisi CO_2 tahunan Indonesia menggunakan data publik dari Our World in Data periode 1990–2022. Tiga variabel prediktor yang digunakan adalah GDP per kapita, konsumsi energi primer, dan populasi. Data dibagi secara kronologis, yaitu 1990–2015 sebagai data latih dan 2016–2022 sebagai data uji. Ridge Regression menggunakan prediktor yang distandardisasi, dan parameter alpha dipilih pada data latih menggunakan TimeSeriesSplit. Kinerja model dievaluasi menggunakan MAE, RMSE, dan R^2, serta dibandingkan dengan baseline Naive Forecast dan Linear Trend. Linear Regression menghasilkan kinerja terbaik pada data uji dengan MAE 19,181, RMSE 21,384, dan R^2 0,887. Ridge Regression menghasilkan MAE 27,693, RMSE 33,758, dan R^2 0,719, namun tetap lebih baik dibandingkan baseline sederhana. Hasil ini menunjukkan bahwa meskipun regularisasi memiliki dasar teoretis pada kondisi* *multikolinearitas, penerapannya pada dataset tahunan yang kecil justru menurunkan akurasi pada data uji. Temuan penelitian ini sebaiknya dibaca sebagai kinerja prediktif, bukan hubungan sebab-akibat.*

Keywords: Emisi CO_2, Regresi Linier, Regresi Ridge, Prediksi.

INTRODUCTION

Monitoring carbon dioxide (CO_2) emissions at the national level has become one of the primary references for assessing the progress of global climate change policies (IPCC, 2023). In the Indonesian context, annual emissions data since the early 1990s reveal an increasing trend that corresponds with growth in economic output, energy consumption, and population (Ritchie et al., 2026; Wahyudi et al., 2024). A quantitative study covering the period 1990–2021 found that GDP per capita has a positive and significant effect on Indonesia's CO_2 emissions, while population exerts a significant long-term influence (Wahyudi et al., 2024). A similar pattern has also been observed across ASEAN countries, where long-run panel evidence identifies per capita income and energy consumption as the two main contributors to regional CO_2 emissions (Ponce & Manlangit, 2023), with Indonesia being one of the largest emitters in the region (Ritchie et al., 2026). Understanding these dynamics provides an important basis for policy formulation, particularly as Indonesia advances toward its 2060 net-zero emissions target, which requires more rigorous empirical monitoring (Resosudarmo et al., 2023).

Predictive approaches provide insights into the future trajectory of carbon emissions (Cai et al., 2024; Gurcan, 2024; Karakurt & Aydin, 2023). For countries whose energy systems remain heavily dependent on fossil fuels, Karakurt & Aydin (2023) demonstrated that simple regression models continue to provide value for emissions forecasting in the BRICS and MINT countries, a group that includes Indonesia. At the global level, Costantini et al. (2024) compared multiplicative regression models with Random Forest across 117 countries and found that both produced satisfactory forecasting performance, although Random Forest achieved higher predictive accuracy. In the context of developing countries, Kumari & Singh (2023) evaluated six forecasting models, ARIMA, SARIMAX, Holt–Winters, Random Forest, LSTM, and Linear Regression, to predict India's CO_2 emissions. Their results showed that LSTM achieved the best performance, while SARIMAX and Holt–Winters also demonstrated high forecasting accuracy. The availability of open-access repositories such as Our World in Data further facilitates similar analyses by providing consistent annual country-level indicators covering economic, energy, and demographic variables (Ritchie et al., 2026).

Linear Regression is widely employed as a starting point for predictive analysis because of its simplicity and the straightforward interpretability of its coefficients (Hastie et al., 2009). Model coefficients are estimated by minimizing the sum of squared differences between the observed and predicted values. However, this method becomes less reliable when predictor variables exhibit similar patterns of variation, as the estimated coefficients become unstable and highly sensitive to small changes in the data (Hastie et al., 2009; Luqman et al., 2025).

Ridge Regression addresses this limitation by introducing an L2 penalty into the objective function. This regularization constrains coefficient magnitudes, thereby producing more stable estimates in the presence of multicollinearity (Hoerl & Kennard, 1970; Magklaras et al., 2024). Recent developments in ridge estimators further reinforce the relevance of this method for datasets with highly correlated predictors (Akhtar et al., 2024; Alharthi & Akhtar, 2025; Shaheen et al., 2023). In macroeconomic datasets, variables such as GDP per capita, primary energy consumption, and population frequently exhibit this characteristic because they tend to increase simultaneously over time.

The application of machine learning methods to CO_2 emission forecasting has been widely explored across various contexts. Gurcan (2024) employed ensemble learning to estimate motor vehicle emissions, while Cai et al. (2024) developed a hybrid LightGBM–XGBoost model for carbon emission estimation based on electricity–carbon relationships. Begum & Al Mobin (2025) projected emissions for the world's eleven largest emitting countries through 2030. A broader comparative study by Ajala et al. (2025) evaluated fourteen statistical, machine learning, and deep learning models using daily emissions data from the four largest emitting regions, demonstrating that machine learning and deep learning models generally outperformed conventional statistical approaches at the daily resolution. At the city level, Lei et al. (2024) found that Random Forest outperformed multiple linear regression using data from 281 Chinese cities, while Foong et al. (2024) reported that a multilayer perceptron (MLP) combined with a nature-inspired optimization algorithm achieved high predictive accuracy for the OECD Asia–Oceania region. In Indonesia, a panel study by Kurniawan et al. (2025) modeled the determinants of CO_2 emissions in West Java, highlighting the roles of population growth, industrial investment, and private vehicle ownership. Unlike these previous studies, the comparison between two linear regression variants in the present study is positioned as a baseline predictive approach, not to outperform more sophisticated models, but to evaluate the extent to which simple regression models can capture the dynamics of Indonesia’s national CO_2 emissions.

This study aims to evaluate the predictive performance of Linear Regression and Ridge Regression as baseline models for forecasting Indonesia’s annual total CO_2 emissions over the period 1990–2022, using three predictors: GDP per capita, primary energy consumption, and population. The performance of both models is compared on the 2016–2022 test dataset using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). To provide a more comprehensive evaluation framework, the study also incorporates two simple benchmark models, namely the Naive Forecast and Linear Trend. The primary focus of this research is prediction rather than causal inference.

METHODS

This study is quantitative in nature. Two main regression models were compared for the task of predicting CO_2 emissions: Linear Regression and Ridge Regression. In addition, two simple baseline models were used as a point of comparison to evaluate the performance of the regression models against simpler predictive approaches. This study does not aim to draw conclusions about the direction of the relationships between variables.

Data Sources and Research Variables

The data are sourced from the openly available OWID CO_2 and Greenhouse Gas Emissions Dataset (Ritchie et al., 2026). This repository contains CO_2 emissions data and related indicators for various countries. The dataset was filtered to include data for Indonesia from 1990 to 2022, this range was selected because all required variables were fully available without any missing values throughout that period. A total of 33 annual observations were used.

Four variables were derived from the dataset: one target variable and three predictor variables. The target variable is Indonesia’s total annual CO_2 emissions. The three predictor variables consist of GDP per capita, primary energy consumption, and population. GDP per capita was not taken directly from the dataset columns but was calculated from the ratio of GDP to population for the same year. Details of the variables are presented in Table 1.

Table 1. Research Variables

Variables	Role	Definition / Calculation / Unit
Total Emission CO_2	Target (Y)	Total emission CO_2 Indonesia; CO2 column; million metric tons/year

Variables	Role	Definition / Calculation / Unit
GDP per Capita	Predictor (X_1)	gdp / population; US Dollar per capita
Primary Energy Consumption	Predictor (X_2)	primary_energy_consumption; TWh/year
Population	Predictor (X_3)	population; people

Source: research data

Table 2 shows five rows from the final dataset to provide an overview of the data format, units, and the training/test split. The complete dataset is provided in a separate Excel file, and the composition of the split is summarized in Table 3.

Table 2. Example of the Structure of Indonesia's Final Dataset for the Period 1990–2022

Year	CO_2 (million metric tons)	Predictor and split summary
1990	154,991	GDP/capita 3.915,77; energy 600,179 TWh; population 183.501.097; Training
1991	174,754	GDP/capita 4.210,56; energy 652,938 TWh; population 186.778.239; Training
1992	199,298	GDP/capita 4.429,40; energy 716,860 TWh; population 190.043.742; Training
2021	632,946	GDP/capita 12.009.24; energy 2.196,681 TWh; population 276.758.056; Testing
2022	758,021	GDP/capita 12.552.76; energy 2.749,242 TWh; population 278.830.527; Testing

Source: research data

Table 3. Training and Testing Data Split

Period	Number of Observations	Usage
1990–2015	26 observations	Training data
2016–2022	7 Observations	Test Data

Data Preprocessing and Data Splitting

Preprocessing was conducted in stages. The initial dataset was filtered to include only Indonesia, then restricted to the 1990–2022 range. The retained columns were adjusted to meet the needs of the analysis: year, co2, gdp, population, and primary_energy_consumption. GDP per capita was then calculated from the ratio of GDP to population. The final step was a recheck to ensure there were no missing values in the variables to be used.

Specifically for Ridge Regression, the predictor variables were standardized using StandardScaler so that differences in units between variables would not interfere with the regularization penalty; this choice is consistent with the findings of Sujon et al. (2024), who showed that the performance of linear models with regularization is sensitive to the scaling technique used. StandardScaler is placed within the pipeline alongside Ridge Regression. Thus, the scaling process is learned solely from the training data during each validation iteration, thereby minimizing the risk of data leakage (Kapoor & Narayanan, 2023). Linear Regression is run directly on the variables' original scales for prediction evaluation, while standardized coefficients are calculated separately to compare the relative contributions of the variables.

The data was divided chronologically without randomization. The 1990–2015 range was designated as the training data, and 2016–2022 as the test data. The temporal nature of the emissions data makes a chronological scheme more appropriate: the model is trained on the historical period and tested on the subsequent period. This approach also prevents future information from leaking into the training process, which is a common source of data leakage in time-series contexts (Kapoor & Narayanan, 2023).

Regression Model and Implementation Workflow

The model was implemented in a sequential series of steps. First, the Indonesian data for the 1990–2022 period was organized, and the variables in Table 1 were prepared. Next, the data is divided chronologically into training and test sets. Linear Regression is trained on the 1990–2015 training data to obtain predictions for 2016–2022. Ridge Regression is trained using the StandardScaler and Ridge pipeline, while the optimal alpha value is selected via GridSearchCV with TimeSeriesSplit on the training data. Two simple baselines were also calculated: Naive Forecast and Linear Trend. After all predictions were obtained, model performance was compared using MAE, RMSE, and R^2.

Linear Regression seeks a combination of coefficients that relates three predictors, GDP per capita, primary energy consumption, and population, to total CO_2 emissions. During the training process for the 1990–2015 period, the coefficients were adjusted so that the sum of the squares of the differences between the predictions and the actual values was minimized (Hastie et al., 2009). The equation is as follows:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{1}$$

where $\hat{y}$ represents the predicted CO_2 emissions, β_0 is the intercept, and β_1, β_2, and β_3 are, respectively, the coefficients for GDP per capita, primary energy consumption, and population. After training is complete, the equation is applied to the 2016–2022 test data to generate annual predictions.

Ridge Regression uses the same framework to estimate coefficients for predicting CO_2 emissions but adds a component to its objective function. This component is an L2 penalty that prevents the coefficients from becoming excessively large when predictors are correlated with one another (Hoerl & Kennard, 1970; Magklaras et al., 2024). The objective function for Ridge Regression is formulated as:

$$\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \alpha \|\boldsymbol{\beta}\|^2 \right\} \tag{2}$$

The first term measures prediction inaccuracy, while the second term imposes a penalty on the magnitude of the coefficients. The parameter α controls the intensity of the penalty: a larger value of α suppresses the magnitude of the coefficients more strongly (Shaheen et al., 2023). The value of α was selected using GridSearchCV across 100 candidate values distributed logarithmically from 10^{-3} to 10^3. The selection of α was performed only on the 1990–2015 training data using `TimeSeriesSplit` with five-fold cross-validation to ensure that the validation process accurately reflected the temporal characteristics of the data. The optimal value of α obtained was 0.61359. Once α was set, the final model was trained on the entire training dataset and used to predict emissions for the 2016–2022 test data. As additional comparators, the Naive Forecast predicts emissions for a given year using the previous year's emission value, while the Linear Trend uses year as the sole predictor.

Model Evaluation

After predictions for 2016–2022 were obtained from the two main models and the two baselines, the results were compared with actual CO_2 emissions for the same years. The difference between the actual and predicted values is called the residual and serves as the basis for calculating MAE, RMSE, and R^2 (Chicco et al., 2021).

MAE calculates the average magnitude of the residuals without distinguishing the direction of the error, so predictions that are higher or lower than the actual values are given equal weight. RMSE also measures the magnitude of the residuals, but imposes a greater penalty on large errors due to the use of the square. R^2 indicates how much of the variation in the test data is explained by the model; a value closer to 1 indicates a better fit (Chicco et al., 2021).

RESULTS AND DISCUSSION

Descriptive Statistics and Correlation

The four variables used, CO_2 emissions, GDP per capita, primary energy consumption, and Indonesia's population for the period 1990–2022, have different units: emissions in millions of metric tons, GDP per capita in dollars, primary energy consumption in TWh, and population in number of people. Because their value ranges differ significantly, the trends are visualized as indices with a base year of 1990 = 100 so that changes in each variable can be compared on a single graph.

Figure 1 presents these trends, with the X-axis showing the year of observation and the Y-axis showing the index values. The indices in the graph are used solely for visualizing trends and are not used as inputs for the regression model; the model is trained using the original values of each variable.

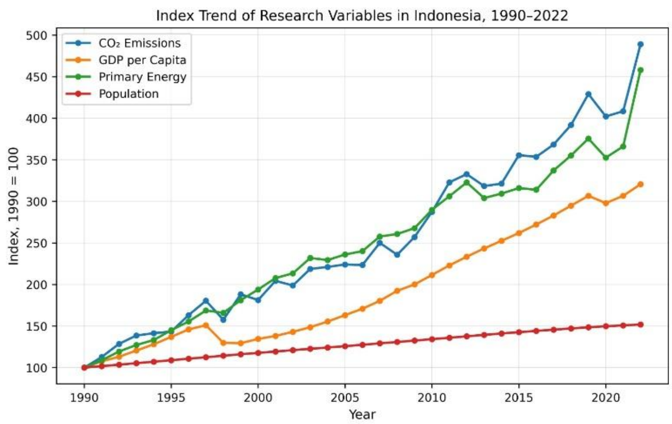

Figure 1. Trend of research variable index for the period 1990–2022 with the base year 1990 = 100

Figure 1 shows a consistent trend: all four variables increased between 1990 and 2022. The sharpest increase, measured from the starting point, occurred in CO_2 emissions and primary energy consumption. GDP per capita followed with a consistent increase, while the population grew at the slowest rate. The uniform direction of change across the variables provides an initial indication of a strong correlation that needs to be considered when selecting a model.

The relationships among the variables were further examined using Pearson's correlation, as shown in Figure 2. Pearson's correlation describes the strength and direction of the linear relationship between two variables on a scale from -1 to 1; values close to ±1 indicate a strong linear relationship, while values close to 0 indicate a weak relationship.

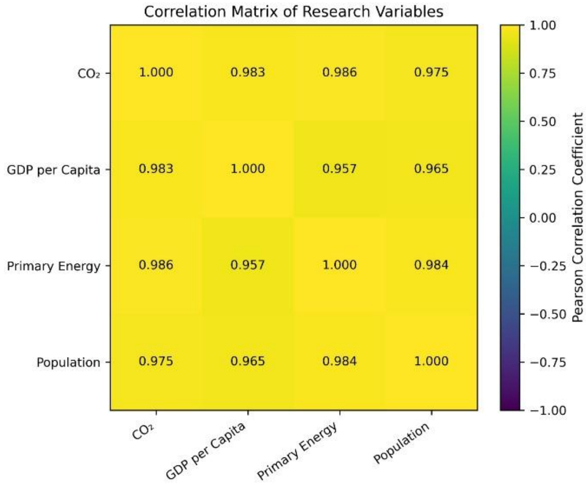

Figure 2. Correlation matrix between research variables

Figure 2 shows that CO_2 emissions are strongly correlated with all three predictors. The correlations among the predictors are also high, meaning that all three predictors follow similar patterns. In linear regression, such conditions can cause the coefficients to become unstable (Hastie et al., 2009; Luqman et al., 2025). For this reason, Ridge Regression was used as a comparison because its L2 regularization is specifically designed to mitigate coefficient instability under multicollinearity (Hoerl & Kennard, 1970; Magklaras et al., 2024). It should be emphasized that high correlation values should not be interpreted as evidence of a cause-and-effect relationship; these figures merely describe the co-occurrence of patterns during the observation period, in line with repeated warnings in the scientific machine learning literature regarding the distinction between predictive associations and causal claims.

The severity of this multicollinearity is clearly evident from the Variance Inflation Factor (VIF) values. Calculated over the entire observation period (33 observations), the VIF values reached approximately 15 for GDP per capita, 31 for primary energy consumption, and 38 for population. These values remained unchanged in both the raw and standardized data, as VIF is not affected by scaling. All three values are well above 10, the threshold commonly used to indicate serious multicollinearity. This finding reinforces the rationale for using Ridge Regression and also helps explain why the population coefficient can be negative, as coefficient estimates become unstable when predictors carry highly overlapping information.

Model Evaluation Results

Table 4 summarizes the performance of Linear Regression, Ridge Regression, and the two baseline models on the 2016–2022 test data. Each model produced seven annual predictions, and the differences between these predictions and the actual values were used to calculate MAE, RMSE, and R^2.

Table 4. Comparison of Model and Baseline Performance on Test Data (2016–2022)

Model	MAE	RMSE	R^2
Naive Forecast	42.390	56.910	0.202
Linear Trend	56.418	69.959	-0.206
Linear Regression	19.181	21.384	0.887
Ridge Regression	27.693	33.758	0.719

Source: research data

Linear Regression performed best on the test data, with an MAE of 19.181, an RMSE of 21.384, and an R^2 of 0.887. Ridge Regression yielded an MAE of 27.693, an RMSE of 33.758, and an R^2 of 0.719. Although Ridge Regression did not outperform Linear Regression, its performance was still better than that of the two simple baselines. Naive Forecast produced an MAE of 42.390, an RMSE of 56.910, and an R^2 of 0.202, while Linear Trend produced an MAE of 56.418, an RMSE of 69.959, and an R^2 of -0.206.

These results indicate that the use of economic, energy, and demographic predictors provides better predictive power compared to simple approaches based solely on the previous year's values or annual trends. However, on this small annual dataset, Ridge regularization did not yield an improvement in accuracy compared to Linear Regression. This difference should be interpreted with caution, as the test data consists of only seven annual observations.

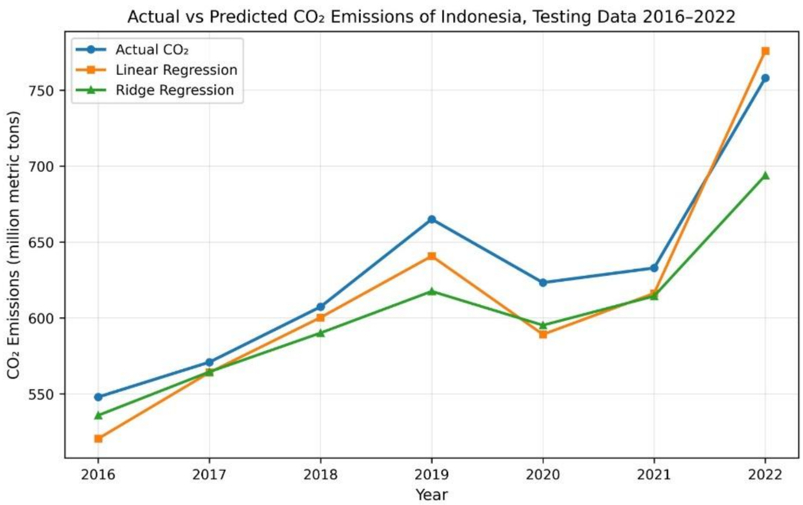

Figure 3. Comparison of actual CO_2 emissions and model predictions in 2016–2022 testing data

Figure 3 shows a comparison between actual CO_2 emissions and the predictions from Linear Regression and Ridge Regression using the 2016–2022 test data. The X-axis represents the year, while the Y-axis represents CO_2 emissions in millions of metric tons. The actual line represents the original data, while the other two lines represent the predictions from each model.

Both models follow the general trend of the test data, but Linear Regression appears closer to the actual values in some years, particularly in 2019 and 2022. Ridge Regression tends to be more conservative due to regularization, resulting in a lower prediction for 2022 than the actual value. This visual pattern aligns with Table 4, where Linear Regression has lower MAE and RMSE values and a higher R^2 compared to Ridge Regression.

For the annual inspection, Table 5 presents the actual values alongside the predictions from both models.

Table 5. Comparison of Actual Values and CO_2 Emission Predictions in Testing Data

Year	Current CO_2 Levels	Linear Regression Predictions	Ridge Regression Prediction
2016	547.969	520.552	535.888
2017	570.928	564.122	564.538
2018	607.378	600.341	590.173
2019	665.049	640.776	617.591
2020	623.304	589.174	595.233
2021	632.946	616.254	614.418
2022	758.021	775.932	693.906

Source: research data

In 2022, the actual emissions were 758.021 million metric tons. Linear Regression predicted 775.932 million metric tons (a difference of +17.911), while Ridge Regression predicted 693.906 million metric tons (a difference of -64.115). For that final year, Linear Regression was closer to the actual value. The magnitude of the Ridge Regression error in 2022 helps explain why the Ridge Regression error metric is higher than that of Linear Regression.

Comparison of Standardized Predictor Coefficients

In addition to error metrics, standardized coefficients are used to examine the relative contribution of each predictor to the prediction. The coefficients in this section are presented on a standardized predictor scale so that the magnitudes of the variables can be compared relative to one another. The results are presented in Table 6.

Table 6. Comparison of Predictor Standardized Coefficients on Both Models

Predictor Variables	Linear Regression	Ridge Regression
GDP per Capita	49.534	42.176
Primary Energy Consumption	111.668	48.711
Population	-49.483	19.615

Source: research data

Primary energy consumption has the largest standardized coefficient in both models; thus, within the models' predictive space, it is the most informative variable among the three predictors. This finding is consistent with the regional literature, which identifies energy consumption as one of the main contributors to CO_2 emissions in ASEAN countries (Ponce & Manlangit, 2023). GDP per capita also makes a substantial predictive contribution. Population has different coefficient signs between Linear Regression and Ridge Regression: negative in Linear Regression, but positive in Ridge Regression. This difference in sign indicates that coefficient interpretation must be approached with caution when predictors are highly correlated. Therefore, the figures in Table 6 should be interpreted as predictive weights within the model, rather than as statements about the direction of empirical effects.

Discussion

The evaluation results show that Linear Regression yields the best predictive performance on the 2016–2022 test data. This finding does not match the theoretical expectation underlying the use of Ridge Regression, which is relevant because GDP per capita, primary energy consumption, and population are highly correlated throughout the observation period. This gap does not invalidate that theoretical basis; rather, it indicates that on an annual dataset with a limited number of observations, stronger regularization does not always improve out-of-sample accuracy. Ridge's lower performance does not indicate that regularization is conceptually irrelevant; rather, it reflects a mismatch between the alpha selection criteria and performance on the test data. Further examination shows that at small alpha values (below 0.1), Ridge performs on par with Linear Regression; performance only declines sharply as alpha increases. The alpha value of 0.61359 selected via TimeSeriesSplit, a procedure that operates on a very limited training subset, over-regularizes the Ridge coefficients, causing the model to become less responsive and to miss the mark when predicting test data. This effect is most evident in the 2022 prediction, which falls far below the actual value.

Primary energy consumption was the variable with the greatest predictive weight in both models. This pattern aligns with the high correlation between primary energy consumption and CO_2 emissions previously observed in Figure 2. The direction of this relationship is also consistent with a panel study in the ASEAN region, which found that increases in energy consumption are associated with nearly proportional increases in CO_2 emissions (Ponce & Manlangit, 2023). Although intuitive, it is important to reiterate that a large predictive contribution does not constitute evidence of a causal relationship. These findings merely indicate that primary energy consumption was the most helpful variable for the models in predicting emissions based on the data structure used.

The difference in the signs of the population coefficients should also be interpreted as a consequence of multicollinearity. In multiple regression, a variable's coefficient indicates its contribution after other variables have been accounted for. Since population, GDP per capita, and primary energy consumption all increase over the long term, the model may distribute predictive information among variables in ways that do not always align with causal intuition. Therefore, this study does not interpret the coefficients as empirical effects but rather as weights in the prediction function.

CONCLUSIONS AND RECOMMENDATIONS

Conclusion

An evaluation of Linear Regression and Ridge Regression as baseline models for predicting Indonesia's CO_2 emissions for the period 1990–2022 shows that Linear Regression performed best on the 2016–2022 test data, with an MAE of 19.181, an RMSE of 21.384, and an R^2 of 0.887. Ridge Regression yielded an MAE of 27.693, an RMSE of 33.758, and an R^2 of 0.719. Although Ridge Regression is theoretically relevant for multicollinearity conditions, the results of this study indicate that regularization does not always improve prediction accuracy on small annual datasets. Both regression models outperformed the simple baselines, namely, the Naive Forecast and Linear Trend, thereby demonstrating that regression models based on economic, energy, and demographic indicators can still serve as an initial baseline for predicting Indonesia's CO_2 emissions. Primary energy consumption emerged as the variable with the greatest predictive weight in both models; however, this finding should be interpreted as a predictive contribution rather than evidence of a causal relationship.

Suggestions for future research. The main limitation of this study is the small sample size, which appears to directly affect the stability of Ridge hyperparameter selection. Therefore, the most relevant direction for further development is to expand the dataset, for example, through a cross-country approach, so that the cross-validation procedure yields more reliable parameter selection. Furthermore, since the initial motivation for using Ridge stemmed from strong multicollinearity among predictors, future research could test other methods designed for similar conditions, such as Elastic Net or Partial Least Squares, as a comparison to these baseline results.

Limitations

This study has several limitations. First, the test data contains only seven annual data points because the data resolution is annual; consequently, performance metrics are sensitive to the behavior of individual data points, and estimates of model stability are not as robust as they would be with a longer dataset. Second, the scope of the analysis is limited to three aggregate predictors and does not include other dimensions such as the energy mix, industrial structure, or climate policies; more comprehensive studies of emission determinants, such as Kurniawan et al. (2025) for West Java, show that other variables are also relevant when available. Third, the 2016–2022 test period includes an unusual disruption in 2020 due to the pandemic, which temporarily suppressed global emissions; the model can still track broad trends but is not designed to explicitly capture external shocks. Fourth, a linear regression with three aggregate predictors is not expected to be on par with more complex machine learning approaches such as Random Forest or gradient boosting (Ajala et al., 2025; Costantini et al., 2024; Foong et al., 2024; Lei et al., 2024). Given these limitations, the results of this study should be viewed as an initial, exploratory, and predictive baseline.

Recommendations

Future studies are encouraged to expand the dataset by including longer observation periods or multi-country data to improve model generalizability and the stability of hyperparameter optimization. Additional explanatory variables, such as renewable energy consumption, industrial activity, urbanization, carbon pricing, or climate policy indicators, should also be considered to better represent the determinants of CO_2 emissions. Furthermore, future research should compare the present baseline models with more advanced machine learning and deep learning approaches, such as Elastic Net, Partial Least Squares, Random Forest, Gradient Boosting, XGBoost, or Long Short-Term Memory (LSTM), to examine whether higher predictive accuracy can be achieved while maintaining model interpretability.

Author Contributions

Conceptualization: I Putu Santrisna, Dewa Gede Hendra Divayana; Methodology: I Putu Santrisna, Dewa Gede Hendra Divayana; Software: I Putu Santrisna; Validation: Dewa Gede Hendra Divayana, I Made Gede Sunarya; Formal Analysis: I Putu Santrisna; Investigation: I Putu Santrisna; Data Curation: I Putu Santrisna; Writing – Original Draft Preparation: I Putu Santrisna; Writing – Review & Editing: Dewa Gede Hendra Divayana, I Made Gede Sunarya; Supervision: Dewa Gede Hendra Divayana, I Made Gede Sunarya. All authors have read and approved the final version of the manuscript.

REFERENCES

Ajala, A. A., Adeoye, O. L., Salami, O. M., & Jimoh, A. Y. (2025). An examination of daily CO2 emissions prediction through a comparative analysis of machine learning, deep learning, and statistical models. *Environmental Science and Pollution Research*, *32*, 2510–2535. https://doi.org/10.1007/s11356-024-35764-8

Akhtar, N., Alharthi, M. F., & Khan, M. S. (2024). Mitigating Multicollinearity in Regression: A Study on Improved Ridge Estimators. *Mathematics*, *12*. https://doi.org/10.3390/math12193027

Alharthi, M. F., & Akhtar, N. (2025). Modified Two-Parameter Ridge Estimators for Enhanced Regression Performance in the Presence of Multicollinearity: Simulations and Medical Data Applications. *Axioms*, *14*, 527. https://doi.org/10.3390/axioms14070527

Begum, A. M., & Al Mobin, M. (2025). A machine learning approach to carbon emissions prediction of the top eleven emitters by 2030 and their prospects for meeting Paris agreement targets. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-04236-5

Cai, Y., Feng, J., Wang, Y., Ding, Y., Hu, Y., & Fang, H. (2024). The Optuna–LightGBM–XGBoost Model: A Novel Approach for Estimating Carbon Emissions Based on the Electricity–Carbon Nexus. *Applied Sciences*, *14*(11). https://doi.org/10.3390/app14114632

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623

Costantini, L., Laio, F., Mariani, M. S., Ridolfi, L., & Sciarra, C. (2024). Forecasting national CO2 emissions worldwide. *Scientific Reports*, *14*. https://doi.org/10.1038/s41598-024-73060-0

Foong, L. K., Blazek, V., Prokop, L., Misak, S., Atamurotov, F., & Khalilpoor, N. (2024). Improve carbon dioxide emission prediction in the Asia and Oceania (OECD): nature-inspired optimisation algorithms versus conventional machine learning. *Engineering*

Applications of Computational Fluid Mechanics, *18*. https://doi.org/10.1080/19942060.2024.2391988

Gurcan, F. (2024). Forecasting CO2 emissions of fuel vehicles for an ecological world using ensemble learning, machine learning, and deep learning models. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/PEERJ-CS.2234

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*, 55–67. https://doi.org/10.1080/00401706.1970.10488634

IPCC. (2023). *Climate Change 2023: Synthesis Report*. https://doi.org/10.59327/IPCC/AR6-9789291691647

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9). https://doi.org/10.1016/j.patter.2023.100804

Karakurt, I., & Aydin, G. (2023). Development of regression models to forecast the CO2 emissions from fossil fuels in the BRICS and MINT countries. *Energy*, *263*. https://doi.org/10.1016/j.energy.2022.125650

Kumari, S., & Singh, S. K. (2023). Machine learning-based time series models for effective CO2 emission prediction in India. *Environmental Science and Pollution Research*, *30*(55), 116601–116616. https://doi.org/10.1007/s11356-022-21723-8

Kurniawan, R., Kautsar, S., Wahyuni, R. N. T., Gio, P. U., Wongsonadi, S. K., Zubaidi, M., & Andayani, F. (2025). Assessing economic and social determinants of carbon emissions towards sustainable development in West Java, Indonesia. *Carbon Research*, *4*, 34. https://doi.org/10.1007/s44246-025-00200-0

Lei, Q., Yu, H., & Lin, Z. (2024). Understanding China's CO2 emission drivers: Insights from random forest analysis and remote sensing data. *Heliyon*, *10*. https://doi.org/10.1016/j.heliyon.2024.e29086

Luqman, M., Bhatti, S. H., Aydin, D., & Jamil, M. (2025). Addressing multicollinearity in general linear model: A novel approach for ridge parameter with performance comparison. *PLoS ONE*, *20*. https://doi.org/10.1371/journal.pone.0335072

Magklaras, A., Gogos, C., Alefragis, P., & Birbas, A. (2024). Enhancing Parameters Tuning of Overlay Models with Ridge Regression: Addressing Multicollinearity in High-Dimensional Data. *Mathematics*, *12*. https://doi.org/10.3390/math12203179

Ponce, B. J. H., & Manlangit, A. T. (2023). Carbon dioxide emissions and the Environmental Kuznets Curve: evidence from the Association of Southeast Asian Nations. *Environmental Science and Pollution Research*, *30*, 100037–100045. https://doi.org/10.1007/s11356-023-29370-3

Resosudarmo, B. P., Rezki, J. F., & Effendi, Y. (2023). Prospects of Energy Transition in Indonesia. *Bulletin of Indonesian Economic Studies*, *59*, 149–177. https://doi.org/10.1080/00074918.2023.2238336

Ritchie, H., Rosado, P., & Roser, M. (2026). *CO2 and Greenhouse Gas Emissions.Our World in Data*. https://github.com/owid/co2-data

Shaheen, N., Shah, I., Almohaimeed, A., Ali, S., & Alqifari, H. N. (2023). Some Modified Ridge Estimators for Handling the Multicollinearity Problem. *Mathematics*, *11*(11). https://doi.org/10.3390/math11112522

Sujon, K. M., Hassan, R. B., Towshi, Z. T., Othman, M. A., Samad, M. A., & Choi, K. (2024). When to Use Standardization and Normalization: Empirical Evidence from Machine

Learning Models and XAI. *IEEE Access*, *12*, 135300–135314. https://doi.org/10.1109/ACCESS.2024.3462434

Wahyudi, H., Gunarto, T., Ciptawaty, U., Aida, N., Yunita, R., & Putri, R. M. (2024). The Influence of Determinats on CO2 Emission in Indonesia for a Decade. *International Journal of Energy Economics and Policy*, *14*, 61–65. https://doi.org/10.32479/ijeep.15132