e-ISSN: 2614-8404
p-ISSN:2776-3234

Prediction of Carbon Emissions in Indonesia Using Machine Learning: A Focus on Environmental Impact

Rizaldi Putra[1*)], Memet Sanjaya[2], Deni Utama[3] ,Berliana[4]
[1234]Program Studi Bisnis Digital Politeknik Takumi
Email: [1]rizaldi.rip@takumi.ac.id, [2]memet.msa@takumi.ac.id, [3]deni.utama@takumi.ac.id,
[4]effendyberliana@gmail.com

Abstract *–* Carbon emissions represent a critical driver of global climate change, exerting profound impacts on environmental sustainability and public health. This research examines Indonesia's carbon emission trends using a comprehensive dataset spanning global emissions from 1960 to 2018, with specific focus on Indonesia, obtained from Kaggle. Employing Linear Regression (LR) as the primary machine learning technique, the study effectively models and forecasts future carbon emission levels for Indonesia. The findings indicate a projected increase in emissions to 2.38 tons per capita annually by 2030, underscoring the urgent need for robust environmental policies.

Keywords: Machine Learning, Carbon Emission, Indonesia, Linear Regression.

I. INTRODUCTION

Numerous studies have demonstrated the significant environmental and health risks associated with increasing carbon emissions. For instance, carbon emissions not only contribute to global warming but also exacerbate public health risks, particularly in rapidly industrializing nations such as China[1]. Indonesia, with its heavy reliance on fossil fuels, faces similar challenges, especially as its emissions continue to rise. A study [2] demonstrates a strong connection between energy consumption and carbon dioxide emissions. An increase in energy use correlates with higher carbon dioxide emissions, primarily due to fossil fuel consumption. Consequently, greater fossil fuel use results in significantly elevated emissions. The environmental consequences, such as rising sea levels, extreme weather events, and water scarcity, further underscore the importance of addressing carbon emissions. According to the United Nations Sustainable Development Goals (SDGs), reducing carbon emissions is essential for achieving environmental sustainability and mitigating the harmful effects of climate change [3].

This study utilizes global carbon emissions data spanning from 1960 to 2018, obtained from publicly accessible open data on Kaggle. This dataset was selected because it provides a comprehensive overview of global carbon emissions, with a specific focus on Indonesia. Indonesia's demographic bonus, projected to peak around 2030, makes this data particularly relevant for developing a predictive model. The objective of this model is to forecast carbon emissions and analyze potential impacts, offering insights that could inform policy and strategic interventions in response to environmental challenges.

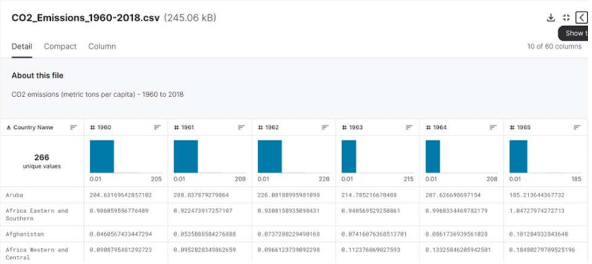
This research utilizes machine learning algorithms - Linear Regression (LR) to predict future carbon emissions in Indonesia. Linear regression remains a foundational method in machine learning due to its simplicity, computational efficiency, and interpretability, particularly when analyzing linear relationships between variables. It is highly valued in applications such as financial forecasting, where transparency in understanding variable influence is essential to ensuring reliability in complex datasets[4]. While models such as support vector machines (SVMs) and multilayer perceptrons (MLPs) may outperform linear regression in handling non-linear or large-scale data, linear regression remains competitive in its niche for clear and efficient predictions. The LR method can produce highly accurate predictive correlation coefficients[5]. The study leverages historical data to provide accurate forecasts that can inform national policies and global efforts to mitigate climate change.

II. RESEARCH METHODOLOGY

This study uses machine learning algorithms, particularly linear regression, to model the relationship between two variables by fitting a linear equation to historical data, with one acting as the explanatory variable and the other as the dependent variable [6], to predict future carbon emissions in Indonesia. The methodology involves data collection, preprocessing, model training and interpretation to generate accurate predictions of future emissions. The analysis is conducted using libraries such as Pandas, NumPy, and Scikit-learn in Python. Libraries like Pandas in Python are among the best available today [7].

A. Data Collection

The data used in this study were sourced from Kaggle, comprising global carbon emissions data from 1960 to 2018, covering 266 countries. The data titled “CO2 emissions (metric tons per capita) from 1960 to 2018" can be access from this link. The dataset was filtered to focus on Indonesia, a country experiencing a significant rise in emissions due to industrial and economic growth. The study indicated that a 1% increase in economic growth and fossil fuel consumption leads to a 0.36% and 0.67% rise in carbon dioxide emissions[8]



CO2_Emissions_1960-2018.csv (245.06 kB)

About this file

CO2 emissions (metric tons per capita) - 1960 to 2018

Country Name	1960	1961	1962	1963	1964	1965
266 unique values	0.01 205	0.01 209	0.01 226	0.01 215	0.01 208	0.01 185
Aruba	204.63169642857102	208.837879279864	226.08188995981098	214.785216670488	207.626698697154	185.213644367732
Africa Eastern and Southern	0.906059556776489	0.922473917257107	0.9308158935098431	0.940569529250861	0.9960334469782179	1.04727974272713
Afghanistan	0.0460567433447294	0.0535888584276888	0.0737208229490168	0.07416076368513701	0.0861736939561028	0.101284932843648
Africa Western and Central	0.0908795481292723	0.0952828349862659	0.0966123739092298	0.112376069027593	0.13325846205942501	0.18480279709525196

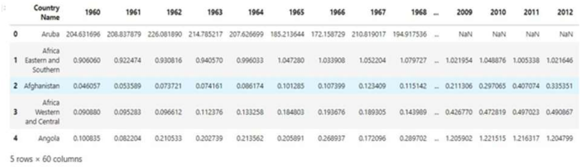
Image 1: carbon emission data from 266 countries

The global carbon emissions dataset from 1960 to 2018 offers a detailed view of CO2 emissions from 266 countries, capturing trends across nearly six decades. It tracks total emissions, per capita figures, and sectoral contributions, revealing a steady rise in emissions, particularly after industrialization, with notable increases from emerging economies in recent decades. The data underscores how economic growth, energy consumption, and population dynamics drive emission patterns, with developed countries contributing heavily in earlier periods and developing nations like Indonesia.

Before training the models, the data underwent preprocessing to ensure quality and consistency. The preprocessing involved the following steps:

1. **Data Import and Cleaning:** Missing values in the dataset were identified and managed through imputation or removal, ensuring that the model would not be biased by incomplete data.

```
df=df.dropna(axis='rows')
df .head()                                   (1)
```



	Country Name	1960	1961	1962	1963	1964	1965	1966	1967	1968	...	2009	2010	2011	2012
0	Aruba	204.631696	208.837879	226.081890	214.785217	207.626699	185.213644	172.158729	210.819017	194.917536	...	NaN	NaN	NaN	NaN
1	Africa Eastern and Southern	0.906060	0.922474	0.930816	0.940570	0.996033	1.047280	1.033908	1.052204	1.079727	...	1.021954	1.048876	1.005338	1.021646
2	Afghanistan	0.046057	0.053589	0.073721	0.074161	0.086174	0.101285	0.107399	0.123409	0.115142	...	0.211306	0.297065	0.407074	0.335351
3	Africa Western and Central	0.090880	0.095283	0.096612	0.112376	0.133258	0.184803	0.193676	0.189305	0.143989	...	0.426770	0.472819	0.497023	0.490867
4	Angola	0.100835	0.082204	0.210533	0.202739	0.213562	0.205891	0.268937	0.172096	0.289702	...	1.205902	1.221515	1.216317	1.204799

5 rows × 60 columns

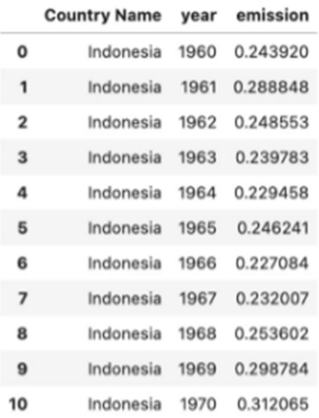
Image 2: Processing data

In data preprocessing, handling missing values is crucial to ensure that the model is not biased or skewed by incomplete data. One common method is removing rows with missing values to maintain the integrity of the dataset. The code provided on Phyton does this by using *df.dropna(axis='rows'),* which removes any rows that contain missing data, and then the *df.head()* function displays the first few rows of the cleaned dataset

2. **Normalization:** Features such as year and emission levels were normalized to standardize the range of data, which is crucial for algorithms sensitive to the scale of input variables.

```
indo=df [df['Country Name']=='Indonesia']
indo                                              (2)

ina = indo.melt(id_vars='Country Name'
     var_name='year,
     value_name='emission'                        (3)
```



	Country Name	year	emission
0	Indonesia	1960	0.243920
1	Indonesia	1961	0.288848
2	Indonesia	1962	0.248553
3	Indonesia	1963	0.239783
4	Indonesia	1964	0.229458
5	Indonesia	1965	0.246241
6	Indonesia	1966	0.227084
7	Indonesia	1967	0.232007
8	Indonesia	1968	0.253602
9	Indonesia	1969	0.298784
10	Indonesia	1970	0.312065

Image 3: Data normalization

Normalization is an essential preprocessing step in machine learning, particularly when using algorithms sensitive to the scale of input variables, such as gradient descent-based models. By normalizing features like the year and emission levels, their values are scaled to a standard range, preventing features with larger ranges from disproportionately influencing the model. This process ensures that each feature contributes equally to the model, improving performance and convergence speed.

The code provided above filters the dataset for Indonesia and reshapes it using the melt function. This operation converts the data from wide format (with columns for each year) to long format, where emission levels are associated with individual years for easier analysis. In this context, the reshaped data allows for more manageable analysis of emission levels over time, facilitating trend identification and model training.

3. **Exploratory Data Analysis:** Visualization techniques were employed to identify trends and patterns in carbon emissions over time. This included plotting historical data to observe fluctuations and averages.

```
import plotly.express as px                       (4)

L = px.scatter (data_frame=ina, x='tahun',
y='emisi', trendline='ols')
```

```
L.show()                                  (5)
```

Exploratory Data Analysis (EDA) is an essential step in understanding the dataset and identifying underlying trends, patterns, and potential relationships. Visualization techniques, such as scatter plots, help in observing the fluctuations, trends, and averages of carbon emissions over time. The code provided uses Plotly Express to create a scatter plot of emissions data for Indonesia, with a trendline fitted using Ordinary Least Squares (OLS) regression. This allows for a clearer view of the historical relationship between year and emissions, highlighting any potential upward or downward trends.

The scatter plot provides a visual representation of carbon emissions over time, while the OLS trendline shows the overall direction of the data, allowing for easy identification of trends and changes. This technique is essential for understanding the impact of various factors on emissions and for making informed predictions

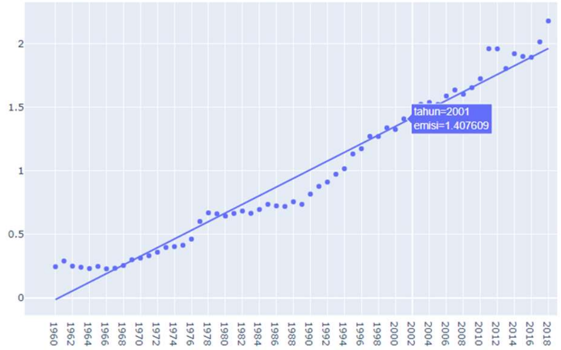


Image 4: Emission trends in Indonesia

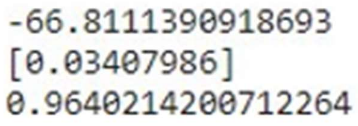
The graph above illustrates a consistent upward trend in carbon emissions over time. The relatively narrow gaps between data points suggest that the rate of growth has remained steady throughout the observed period, indicating a linear progression. This pattern supports the use of linear regression as an appropriate method for creating prediction models, as the method is particularly effective when data exhibits a clear, linear relationship between the independent and dependent variables. Linear regression's simplicity and efficiency make it well-suited for modeling scenarios where growth patterns follow a steady, predictable trajectory

B. Model Training and Interpretation

The algorithm was trained on the historical data to establish a linear relationship between the independent variable (year) and the dependent variable (carbon emissions). The performance of the model was evaluated using the R-squared metric value ranges from 0 to 1, with higher values indicating that the model better explains the variation in the data. The use of R-square provides consistent insights without confusing interpretations, making it a reliable choice for evaluating regression analyses in various fields (Chicco et al., 2021).

```
model= LinearRegression().fit(x,y)
a=model.intercept_
b=model.coef_
r=model.score(x,y)
print(a)
print(b)
print(r)                                          (6)
```

Model training involves fitting an algorithm to historical data to establish a relationship between the independent variable (year) and the dependent variable (carbon emissions). In this case, the linear regression model is trained using the *LinearRegression().fit(x, y)* function, where x represents the year and y represents the emissions. The model's performance is then evaluated using the R-squared metric, which quantifies how well the independent variable(s) explain the variation in the dependent variable. An R-squared value ranges from 0 to 1, with higher values indicating better explanatory power of the model. The model's coefficients *(model.coef_)* and intercept *(model.intercept_)* are also retrieved to understand the relationship, where *model.score(x, y)* provides the R-squared value, offering a clear and interpretable measure of the model's accuracy.



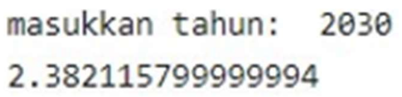
```
-66.8111390918693
[0.03407986]
0.9640214200712264
```

Image 5: result of the data accuracy

Predictions are made using the linear regression formula, where the intercept and coefficients are derived from a previously trained linear regression model. According to the image above, the prediction accuracy is 96%.

```
value=int(input(masukan tahun: '))
prediction=-66.8+(0.03407986*value)
print (prediction)
          (6)
```

The code above allows users to input a year, which is then used in a linear regression formula to predict carbon emissions. The *int(input())* function collects the user input and converts it to an integer, which is stored in the variable nilai. The formula *prediction = -66.8 + (0.03407986 * value)* calculates the emission prediction using the regression model's intercept and coefficient values, which are based on historical data. The predicted value is then displayed using *print(prediction)*. This method provides a simple way to forecast emissions based on the year, leveraging the relationship between time and emissions in a linear regression model



```
masukkan tahun:  2030
2.382115799999994
```

Image 6: result of prediction

Based on the image above, the prediction model estimates that carbon emissions in Indonesia will reach 2.38 units in the year 2030, based on the linear regression formula. This result indicates a consistent upward trend in emissions, aligning with projections of continued growth. By using the formula derived from historical data, the model provides insights into future emissions, allowing policymakers to anticipate environmental impacts and potentially take preventive measures.

We selected the year 2030 for our carbon emissions prediction due to Indonesia's expected demographic bonus in 2045. This five-year gap between 2030 and 2045 is significant point for assessing the potential impacts on various sectors, including the environment and public health. A significant increase in the working-age population is anticipated, which could lead to greater industrial activity, energy consumption, and, consequently, higher carbon emissions. This has a direct impact on public health, as increased emissions can exacerbate air pollution and contribute to respiratory diseases, heat-related illnesses, and other health challenges. Therefore, predicting carbon emissions for this year is crucial for anticipating future health risks and implementing effective policies for sustainable development.

III. RESULTS AND DISCUSSION

A. Prediction Result

To better understanding the prediction result, the data is presented in table below.

Tabel 1: Prediction result from 2025 to 2030

Year	Predicted Emissions (tons per capita)
2025	2.21
2026	2.24
2027	2.27
2028	2.31
2029	2.34
2030	2.38

The prediction models yielded accurate results. The Linear Regression model predicted that Indonesia's carbon emissions will rise to 2.38 tons per capita annually by 2030. The predicted carbon emissions for Indonesia show a steady increase from 2025 to 2030, with emissions per capita rising from 2.21 tons in 2025 to 2.38 tons in 2030. This gradual increase suggests a continued upward trend in emissions as industrial activities and population growth drive energy consumption. These predictions highlight the potential environmental challenges for Indonesia. Effective strategies to manage emissions are essential to mitigate the health risks associated with higher carbon levels, including respiratory issues and other pollution-related diseases.

The projected rise in carbon emissions aligns with global trends, where increasing industrialization drives emissions upward. For instance, previous research [9] demonstrates the effectiveness of machine learning in predicting embodied carbon emissions in residential buildings, reinforcing the utility of predictive models across various contexts. Similarly, studies on ammonia emissions from composting highlight the potential of machine learning in environmental modeling [10].

The findings are consistent with earlier studies that found higher carbon emissions in industrialized countries correlate with public health impacts due to poor air quality [1]. Li et al (2017)[11] Also emphasized the role of household carbon emissions in contributing to indoor and outdoor air pollution.

B. Implication for Policy

The findings of this study highlight the urgent need for Indonesia to implement strong policy measures to manage its carbon emissions. Failure to do so could result in severe environmental degradation, including rising sea levels, biodiversity loss, and public health crises [12]. The healthcare sector also contributes significantly to carbon emissions, underscoring the need for cross-sectoral strategies to reduce environmental damage[13].

Additionally, there is a growing body of evidence suggesting that carbon emissions influence broader economic indicators, such as country risk. (Chaudhry et al., 2020) demonstrated that higher emissions increase the risk profile of G7 economies[12]. Other research[14] discuss the paradox in which economic growth, driven by increased consumption, correlates with lower levels of well-being unless balanced by sustainability measures.

Other studies have also explored machine learning's role in predicting emissions across various sectors. For example, integrating building information modeling (BIM) and machine learning can help predict carbon emissions during construction phases [15]. This potential integration of technology into emissions tracking can provide valuable tools for policymakers and industries aiming to reduce their carbon footprints.

IV. CONCLUSION

This study demonstrates the utility of machine learning algorithms—specifically Linear Regression in predicting future carbon emissions in Indonesia. The model provides accurate forecasts, projecting that Indonesia's carbon emissions will reach 2.38 tons per capita annually by 2030. These findings underscore the urgent need for robust environmental policies aimed at reducing emissions and addressing the environmental challenges posed by rapid industrialization and urbanization.

The predicted increase in carbon emissions highlights potential environmental degradation and raises significant public health concerns. Policymakers must take proactive measures to mitigate emissions across various sectors, including energy, transportation, and healthcare. Future research could expand on this study by incorporating additional factors such as deforestation rates, industrial output, and energy consumption patterns, thereby refining predictive models and enhancing the effectiveness of environmental strategies. By leveraging machine learning and data analysis, Indonesia can develop informed policies that prioritize sustainability and public health, ultimately contributing to global efforts to combat climate change.

References

[1] H. Dong, M. Xue, Y. Xiao, and Y. Liu, “Do carbon emissions impact the health of residents? Considering China’s industrialization and urbanization,” *Sci. Total Environ.*, 2021, doi: 10.1016/j.scitotenv.2020.143688.

[2] E. Lismiyah, M. Marselina, A. R. Taher, T. Gunarto, and N. Aida, “The Causality Between Energy Consumption and Carbon Emission in Indonesia,” *J. Ris. Ilmu Ekon.*, vol. 4, no. 1, pp. 27–38, 2024, doi: 10.23969/jrie.v4i1.83.

[3] E. Rehfuess, S. Mehta, and A. Prüss-Üstün, “Assessing household solid fuel use: Multiple implications for the Millennium Development Goals,” *Environ. Health Perspect.*, 2006, doi: 10.1289/ehp.8603.

[4] H. Wickham and G. Grolemund, *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data*. 2023.

[5] T. Setiyorini and R. T. Asmono, “Komparasi Metode Neural Network, Support Vector Machine Dan Linear Regression Pada Estimasi Kuat Tekan Beton,” *J. TECHNO Nusa Mandiri*, vol. 15, no. 1, p. 51, 2018, [Online]. Available: http://nusamandiri.ac.id

[6] K. Kumari and S. Yadav, “Linear regression analysis study,” *J. Pract. Cardiovasc. Sci.*, 2018, doi: 10.4103/jpcs.jpcs_8_18.

[7] W McKinney, “pandas: a foundational Python library for data analysis and statistics,” *Python high Perform. Sci. Comput.* , 2011.

[8] A. Raihan, D. A. Muhtasim, M. I. Pavel, O. Faruk, and M. Rahman, “An econometric analysis of the potential emission reduction components in Indonesia,” *Clean. Prod. Lett.*, 2022, doi: 10.1016/j.clpl.2022.100008.

[9] X. Zhang, H. Chen, J. Sun, and X. Zhang, “Predictive models of embodied carbon emissions in building design phases: Machine learning approaches based on residential buildings in China,” *Build. Environ.*, vol. 258, 2024.

[10] B. Wang, P. Zhang, X. Qi, G. Li, and J. Zhang, “Predicting ammonia emissions and global warming potential in composting by machine learning,” *Bioresour. Technol.*, vol. 411, 2024.

[11] Q. Li *et al.*, “Impacts of household coal and biomass combustion on indoor and ambient air quality in China: Current status and implication,” *Sci. Total Environ.*, 2017, doi: 10.1016/j.scitotenv.2016.10.080.

[12] S. M. Chaudhry, R. Ahmed, M. Shafiullah, and T. L. Duc Huynh, “The impact of carbon emissions on country risk: Evidence from the G7 economies,” *J. Environ. Manage.*, 2020, doi: 10.1016/j.jenvman.2020.110533.

[13] D. Mominkhan *et al.*, “The current state and potential evolution of carbon emissions in the healthcare sector: a narrative review article,” *Front. Sustain. Energy Policy*, 2023, doi: 10.3389/fsuep.2023.1230253.

[14] A. L. Fanning and D. W. O’Neill, “The Wellbeing–Consumption paradox: Happiness, health, income, and carbon emissions in growing versus non-growing economies,” *J. Clean. Prod.*, 2019, doi: 10.1016/j.jclepro.2018.11.223.

[15] H. Wang *et al.*, “Integrating BIM and machine learning to predict carbon emissions under foundation materialization stage: Case study of China’s 35 public buildings,” *Front. Archit. Res.*, 2024, doi: 10.1016/j.foar.2024.02.008.