Analysis of Individual CO2 Carbon Footprint Predictions in the Environment by Comparing XGBoost Classifier and Cat Boost Classifier Algorithms

Maulisa Syahputri, Zulham Sitorus, Muhammad Iqbal

Abstract
Weather changes caused by global warming have become a global issue that concerns many parties. One of the main factors is greenhouse gases, which trigger the greenhouse effect in the Earth's atmosphere. Currently, carbon dioxide concentrations in the atmosphere are estimated to be at their highest level ever. Carbon emissions originate from organizations, events, products, and human activities, and are referred to as carbon footprints. Carbon footprints serve as indicators of human activities that affect the environment. Carbon emission issues will form a trend from year to year, especially in developing and developed countries that have high motor vehicle consumption. Predictions of individual CO2 carbon footprints on the environment by comparing the xgboost classifier and cat boost classifier algorithms show that the xgboost Classifier is the best performing model with an Accuracy Score of 0.997, followed by the Cat Boost Classifier. This shows that electricity consumption and waste increase the carbon footprint, while renewable energy reduces the carbon footprint and eco actions have an impact but are still small.

Keywords: Carbon Footprint, XGBoost, Cat Boost

Maulisa Syahputri[1]
[1]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: maulisa.syahputri@gmail.com[1]

Zulham Sitorus[2], Muhammad Iqbal[3]
[2,3]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: zulhamsitorus@dosen.pancabudi.ac.id[2], muhammadiqbal@dosen.pancabudi.ac.id[3]
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era
https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Introduction

Climate change caused by global warming is one of the global issues that are of concern to many parties. The main factor that contributes to this is greenhouse gases (GHGs) which trigger the Greenhouse Effect in the Earth's atmosphere[1]-[2]. Currently, carbon dioxide concentrations are estimated to be the highest in the atmosphere. One of the carbon emissions comes from an organization, event, product, and human activity called a carbon footprint. The carbon footprint serves as an indicator of human activities that affect the environment. With the increasing number of activities carried out by humans, the value of the emissions produced will also increase. This shows a link between human activities and air quality in the atmosphere. An important first step for institutions in global warming mitigation efforts is to evaluate greenhouse gas emissions resulting from individual activities, both direct and indirect. This aims to find out how much CO_2 emissions arise from use[3].

Climate change mitigation is a problem that must be addressed to reduce the carbon footprint in the energy sector, which is a question that must be answered. The transition from traditional to renewable energy is often hampered by a variety of technical, financial, and policy issues. How do we address these challenges, and how much can we maximize opportunities to develop clean energy technologies? This study aims to evaluate by predicting the carbon footprint in energy production by identifying the main sources of emissions and analyzing the efficiency of the production process. In addition, the analysis of the carbon footprint in energy use is carried out by understanding energy consumption patterns, their impact on greenhouse gas emissions, and potential reductions at the consumer and industrial sector levels[4].

There have been several previous studies using forecasting, one of which was conducted by Syifa et al. in 2022 to forecast carbon emissions using the SARIMA and LSTM methods. This prediction results in more optimal LSTM when used to forecast carbon emission levels compared to the SARIMA method[5]. Another study using the Neural Network model conducted by Niken in 2023 in his research produced an accuracy of more than 98% with a small error value with five times the error [6]. The next study using forecasting was carried out by Kevin in 2024 using the Linear Regression algorithm to measure the level of CO2 expenditure that produces the best performing type of diesel fuel, with an MSE value of 0.5621[7].

From the above research, it was found that the problem of carbon emissions will form a trend from year to year, especially in developing and developed countries that have high consumption of motor vehicles. However, in reality, the prediction process sometimes encounters obstacles due to a lack of data. Therefore, in this study, the prediction of the Individual Carbon Footprint of CO2 in the Environment was carried out by comparing the XGBoost Classifier and Cat Boost Classifier algorithms. XGBoost is a machine learning method used to solve regression and classification problems using the Gradient Boosting Decision Tree, where each tree is strengthened by the previous tree and the next tree is dependent on each other[8]. CatBoost (Categorical Boosting) is a Gradient Boosted Decision Tree (GBDT)-based algorithm developed to handle categorical data efficiently without the need for one-hot encoding[9].

Literature Review

Global climate change is one of the main issues that has received wide attention from various circles[10]. One of the main causes is the increase in greenhouse gas emissions, especially carbon dioxide (CO_2), which come from human activities such as transportation, industry, and energy consumption[11]. These emissions are known as carbon footprints, which are used as indicators to measure the impact of human activities on the environment.

Several previous studies have addressed the importance of reducing carbon footprints through a variety of approaches, including the use of renewable energy and changing people's behavior.

In addition, technological developments allow the analysis and prediction of carbon footprints using machine learning-based methods[12].

Algorithms such as XGBoost and CatBoost are widely used classification methods because they have high performance in processing complex data. XGBoost is known to improve accuracy through efficient boosting techniques[13], while CatBoost excels at handling categorical data without the need for complex encoding processes . Previous research has shown that these two algorithms are effective in a wide range of prediction cases, including in environmental analysis.

Thus, this study uses the XGBoost and CatBoost algorithms to predict individual carbon footprints, to find out the factors that most influence the increase in carbon emissions and provide the basis for more environmentally friendly decision-making.

Research Methodology

The following are the stages of research used, which are approaches with the type of quantitative research and predictive methods[14]. The following are the stages carried out in the research.

1. Problem Identification
   This research began with the identification of problems related to the increase in carbon footprint which is one of the main causes of greenhouses. Therefore, a system that is able to accurately predict an individual's carbon footprint is needed. The purpose of this study is to implement a comparison of the XGBoost Classifier and Cat Boost Classifier algorithms.
2. Data Acquisition
   The next stage is carried out by collecting data that is the basis for the modeling process. This study uses data from the Central Statistics Agency (BPS), an Indonesian government agency tasked with collecting, processing, and presenting national statistical data, which contains information about individual carbon footprints, including transportation choices, energy use, food habits, waste generation and others related to carbon footprint.
3. Data Preprocessing
   After the data is obtained, the pre-processing stage of data is carried out with the aim of ensuring that the data used is in a clean condition and ready to be processed. The process involves handling missing values, normalizing or standardizing numerical values if necessary, and converting categorical data into numerical forms using encoding techniques. In addition, data is divided into training data and test data with a certain ratio to ensure that the training and testing process of the model runs optimally
4. Model Training
   At this stage, model training was carried out using the XGBoost Classifier and Cat Boost Classifier algorithms, which are one of the algorithms that have high data classification capabilities. Model training is carried out on training data that has been prepared through the pre-processing stage. In an effort to improve model performance, parameter adjustments and cross-validation were carried out to prevent overfitting and strengthen the model's ability to generalize to new data.
5. Feature Importance Analysis
   This stage aims to find out which features contribute the most to the prediction results. The XGBoost Classifier and Cat Boost Classifier have features to calculate the importance of each variable. The results of this analysis provide useful information to understand the main factors that affect the carbot footprint and can be used as a consideration in decision-making.
6. Drawing Conclusions and Suggestions
   The last stage includes drawing conclusions based on the results of model evaluation and feature analysis that has been carried out. This conclusion summarizes how well the XGBoost Classifier and Cat Boost Classifier models predict individual carbon footprints

and what features are most influential. This research also provides suggestions that the developed model can be integrated into decision support systems in the field of environment to help.

Results

The stages in this study include several steps, namely data preprocessing, data, model training, and evaluation of results using a confusion matrix to determine the most optimal model. The first step is to import the required data. This study uses the Python programming language with libraries such as sklearn, numpy, pandas, matplotlib, seaborn and plotly. The dataset used in CSV format.

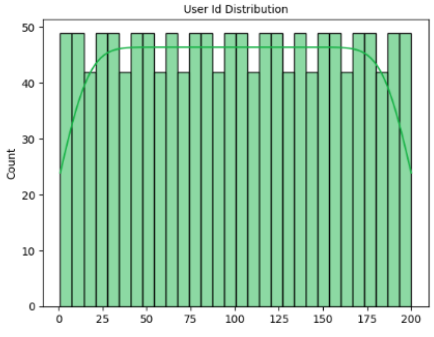
	user_id	day_type	transport_mode	distance_km	electricity_kwh	renewable_usage_pct	food_type	screen_time_hours	wast
0	1	Weekend	EV	1.55	6.12	0	Non-Veg	2.4	0.70
1	1	Weekend	Walk	10.04	4.50	0	Mixed	4.1	0.54
2	1	Weekday	Walk	15.27	2.81	0	Mixed	4.0	0.51
3	1	Weekend	Walk	0.50	10.16	0	Mixed	6.3	0.73
4	1	Weekend	Walk	3.60	5.02	50	Mixed	5.1	0.64
5	1	Weekday	EV	7.77	5.14	50	Veg	4.2	0.73
6	1	Weekend	Bike	6.87	7.14	75	Mixed	2.6	0.45
7	2	Weekend	Bus	0.50	4.76	25	Non-Veg	3.0	0.66
8	2	Weekday	Walk	14.33	8.78	50	Non-Veg	7.4	0.80
9	2	Weekday	Car	9.51	7.78	0	Mixed	2.4	1.03

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1400 entries, 0 to 1399
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   user_id              1400 non-null   int64
 1   day_type             1400 non-null   object
 2   transport_mode       1400 non-null   object
 3   distance_km          1400 non-null   float64
 4   electricity_kwh      1400 non-null   float64
 5   renewable_usage_pct  1400 non-null   int64
 6   food_type            1400 non-null   object
 7   screen_time_hours    1400 non-null   float64
 8   waste_generated_kg   1400 non-null   float64
 9   eco_actions          1400 non-null   int64
 10  carbon_footprint_kg  1400 non-null   float64
 11  carbon_impact_level  1400 non-null   object
dtypes: float64(5), int64(3), object(4)
memory usage: 131.4+ KB
None
```

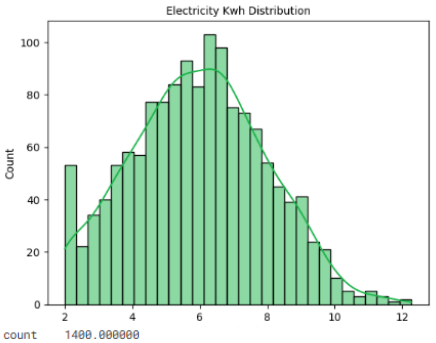
Figure 1. Carbon Footprint Dataset

From the imported dataset , data is displayed as many as 1400 rows and 12 columns consisting of user id, day type, transport model, distance km, electricity kwh, renewable usage, food type, screen time, waste generated, eco action, carbon footprint and canbon impact. Next, it displays the dataset of each variable in the form of a histrogram graph. The following are the results of the carbon footprint histrogram graph.

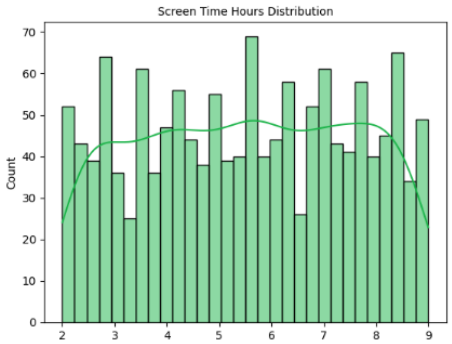
```
count    1400.000000
mean      100.500000
std        57.754936
min         1.000000
25%        50.750000
50%       100.500000
75%       150.250000
max       200.000000
Name: user_id, dtype: float64
```



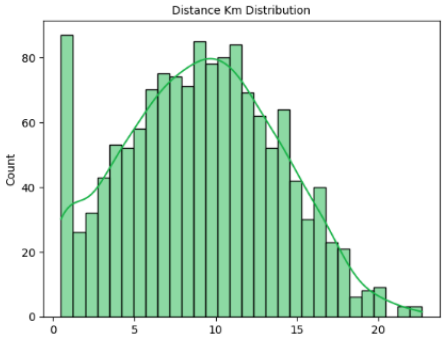

```
count    1400.000000
mean        5.951443
std         1.993266
min         2.000000
25%         4.530000
50%         5.925000
75%         7.290000
max        12.270000
Name: electricity_kwh, dtype: float64
```



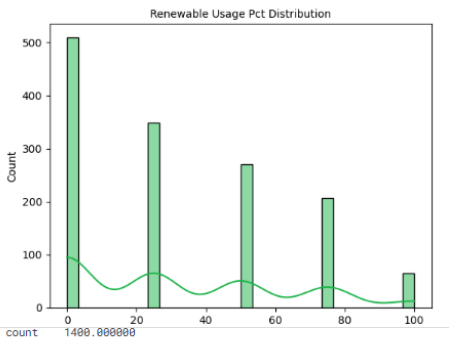

```
count    1400.000000
mean        5.521786
std         2.018918
min         2.000000
25%         3.800000
50%         5.550000
75%         7.300000
max         9.000000
Name: screen_time_hours, dtype: float64
```



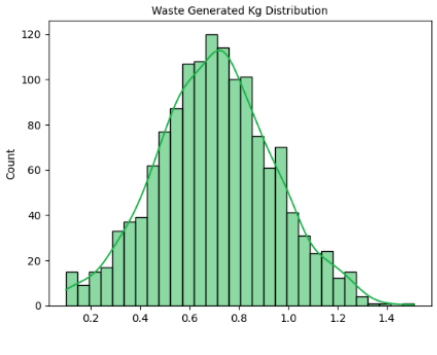

```
count    1400.000000
mean        9.106071
std         4.734692
min         0.500000
25%         5.662500
50%         9.180000
75%        12.510000
max        22.670000
Name: distance_km, dtype: float64
```




```
count    1400.000000
mean       31.589286
std        30.598496
min         0.000000
25%         0.000000
50%        25.000000
75%        50.000000
max       100.000000
Name: renewable_usage_pct, dtype: float64
```

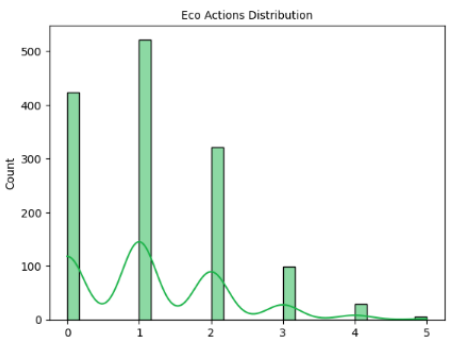



```
count    1400.000000
mean        0.703200
std         0.236415
min         0.100000
25%         0.550000
50%         0.705000
75%         0.860000
max         1.510000
Name: waste_generated_kg, dtype: float64
```



```
count    1400.000000
mean        1.144286
std         1.019066
min         0.000000
25%         0.000000
50%         1.000000
75%         2.000000
max         5.000000
Name: eco_actions, dtype: float64
```

```
count    1400.000000
mean        7.980236
std         2.657858
min         1.790000
25%         6.080000
50%         7.835000
75%         9.830000
max        16.020000
Name: carbon_footprint_kg, dtype: float64
```



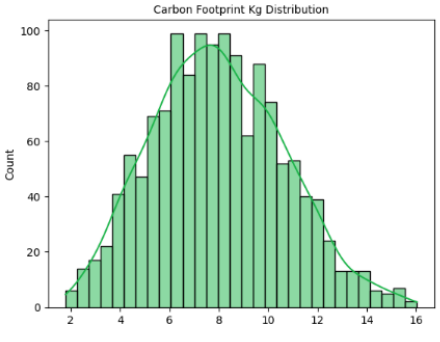


Figure 2. Carbon Footprint Histrogram Graph

From the image above shows the results of the histogram graph of each variable, each variable has a count, mean, std, min, 25%, 50%, 75% and max value. The next stage is to correlate between variables. The following are the results of the correlation between variables.

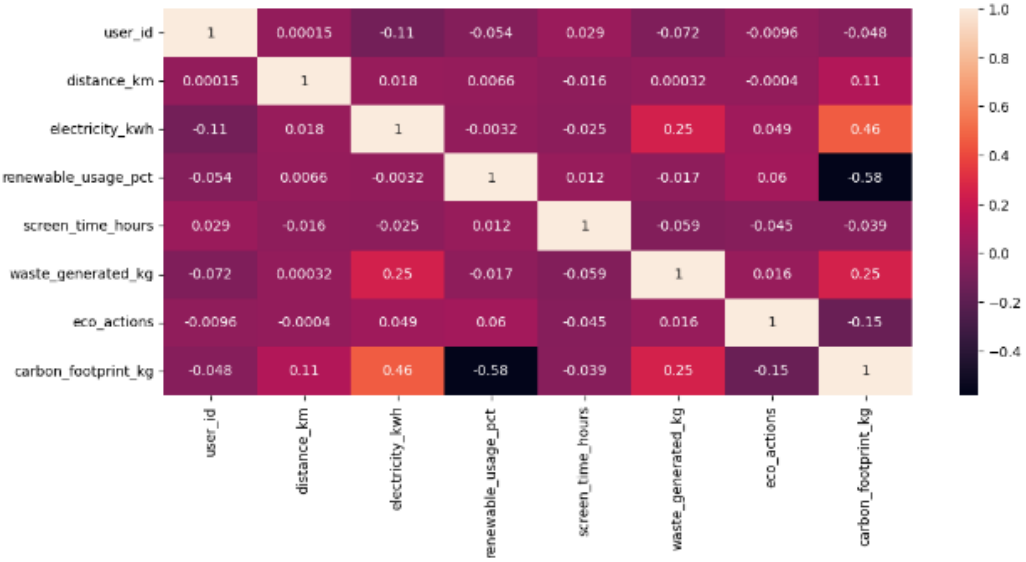


Figure 3. Correlation of Relationships Between Variables

From the image above, it shows that electricity consumption and waste increase the level of carbon footprint, while renewable energy reduces the carbon footprint and eco actions have an impact but are still small. Furthermore, the results of the correlation are visualized in the form of a scatter plot regression graph.

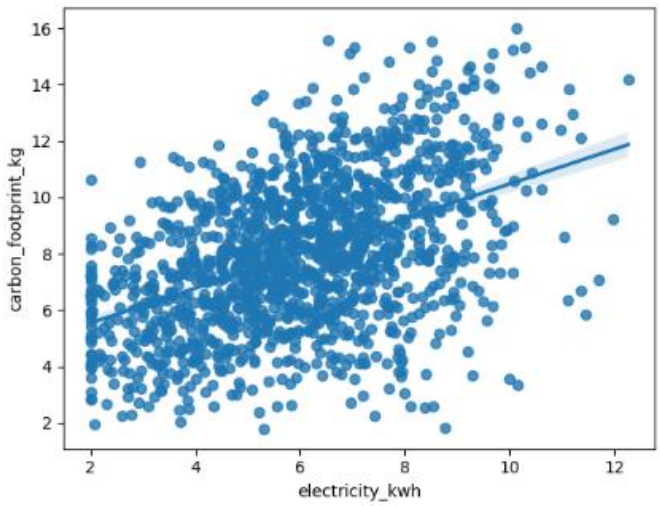
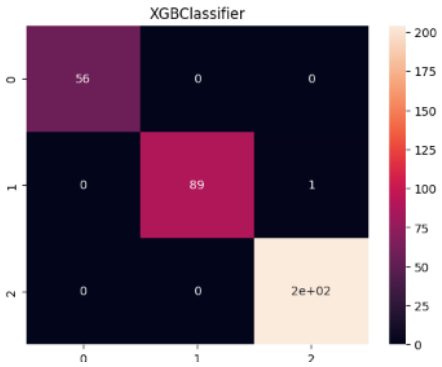


Figure 4. Scatter Plot Regression Chart.

The results of the image above show a positive relationship between electricity consumption and carbon footprint. Increased electricity consumption is likely to be followed by increased carbon emissions, although the relationship is moderate and influenced by other factors. The next stage is to compare the boost algorithms, namely XGBoost and CatBoost. The following are the results of the comparison of the two algorithms.

a. XGBoost Classifier Algorithm



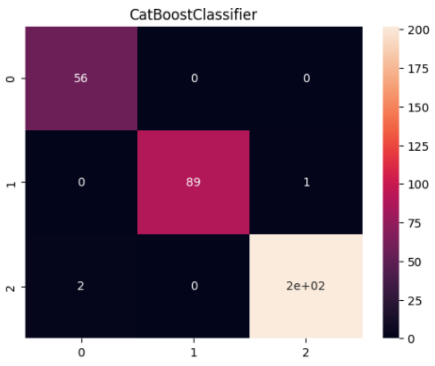
F1 Score:0.9971383691365662

XGBClassifier:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56
1	1.00	0.99	0.99	90
2	1.00	1.00	1.00	204
accuracy			1.00	350
macro avg	1.00	1.00	1.00	350
weighted avg	1.00	1.00	1.00	350

Figure 5. XGBoost Clasifier Algorithm Results

b. Cat Boost Clasifier Algorithm



F1 Score:0.9914601858853549

CatBoostClassifier:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	56
1	1.00	0.99	0.99	90
2	1.00	0.99	0.99	204
accuracy			0.99	350
macro avg	0.99	0.99	0.99	350
weighted avg	0.99	0.99	0.99	350

Figure 6. Results of the Cat Boost Clasifier Algorithm

Table 1. Algorithm Comparison Results

Yes	Models	F1_score	Accuracy_Score	Precision_Score	Recall_Score
1	XGBClassifier	0.997138	0.997143	0.997157	0.997143
2	CatBoostClassifier	0.991460	0.991429	0.991612	0.991429

The results of the comparison table of the XGBoost Clasifier and Cat Boost Clasifier algorithms above show that the accuracy level of the XGBoost Clasifier algorithm is more dominant than the Cat Boost Clasifier algorithm.

Conclusion

The study compared the boost algorithms, namely XGBoost Clasifier and CatBoost Clasifier in predicting the carbon footprint of individual CO2. The test results showed that the XGBoost Clasifier was the model with the best performance with a F1_score value of 0.997, Accuracy_Score : 0.997 and followed by the Cat Boost Clasifier with a value of F1_score : 0.991, Accuracy_Score : 0.991. With the results of electricity consumption and waste increasing the level of carbon footprint, while renewable energy reduces the carbon footprint and eco actions have an impact but are still small. Increased electricity consumption tends to be followed by increased carbon emissions, although the relationship is moderate and influenced by other factors.

References

[1]. Erlindra, D., Febrion, C., Novia, F., Kebangsaan, U., & Indonesia, R. (2024). ANALYSIS OF THE CARBON FOOTPRINT OF THE ENERGY SECTOR FROM CAMPUS ACTIVITIES AT THE NATIONAL UNIVERSITY OF THE REPUBLIC OF INDONESIA, BANDUNG, WEST JAVA. 7, 98–102.

[2]. Rahayuningsih, M., Handayani, L., Abdullah, M., Arifin, M. S., & Kunci, K. (2021). i j. 10(1), 48–52. https://doi.org/10.15294/ijc.v10i1.30038

[3]. Admaja, W. K., Sriwinarno, H., Yogyakarta, T., & Karbon, J. (2018). IDENTIFICATION AND ANALYSIS OF CARBON FOOTPRINT FROM ELECTRICITY USE IN YOGYAKARTA INSTITUTE OF TECHNOLOGY A. 18(2).

[4]. Syah, N., Razak, A., & Diliafrosa, S. (2024). Gudang Jurnal Multidisciplinary Sciences Carbon Footprint Analysis in Energy Production and Consumption: Towards a Green Economy. Consider using the Doctrine of Reason, 197–201.

[5]. Ilma, S., Suwandi, N., Tyasnurita, R., & Muhayat, H. (2022). Carbon Emission Forecasting Using SARIMA and LSTM Methods. 6(1), 73–80.

[6]. World, M., & Data, I. (2023). ALTERNATIVE VEHICLES USE MACHINE LEARNING WITH. 615–625.

[7]. Asgaryansyah, K., Elektro, J. T., & Mataram, U. (2024). Implementation of Linear Regression Algorithm to Measure the Level of Co2 Expenditure in Motor Vehicles. 2(3).

[8]. Asgaryansyah, K., Elektro, J. T., & Mataram, U. (2024). Implementation of Linear Regression Algorithm to Measure the Level of Co2 Expenditure in Motor Vehicles. 2(3).

[9]. Saputra, G. E., Hanindia, M., Swari, P., & Nurlaili, A. L. (2025). Implementation of XGBoost, CatBoost, and LGBM Algorithms for Air Pollution Classification. Scott, 14135–14139.

[10]. Z. Sitorus, A. Putera Utama Siahaan, B. Sugito, A. Ofta Sari, and A. Ibezato Zalukhu, "Digital Marketing Strategy with SEO (Search Engine Optimization) Method for MSMEs in Klambir 5 Garden Village," Journal of Gemilang Community Service (JPMG), vol. 4, 2024, doi: 10.58369/jpmg.v2i4.157.

[11]. Z. Sitorus, E. Hariyanto, and F. Kurniawan, "Analysis of Artificial Intelligence Machine Learning Technology for Mapping and Predicting Flood Locations in Pahlawan Batu Bara Village," 2023.

[12]. A. P. U. Siahaan, M. Iqbal, D. Dika, and M. Syahputri, "Classification Of Pistachio Varieties Using Machine Learning Algorithms," Journal Minfo Polgan, vol. 14, no. 1, pp. 1452–1457, Jul. 2025, doi: 10.33395/jmp.v14i1.15088.

[13]. M. I. Sarif, R. Marbun, M. Syahputri, and J. R. Sitompul, "Digital Empowerment to Support Agricultural MSMEs in Rural Crop Marketing Using the "iConnect Cricket"

App," Jurnal Abdimas TGD, vol. 6, no. 1, January 2026, Page 49-55, P-ISSN : 2809-7289, E-ISSN : 2809-6126, DOI: https://doi.org/10.53513/abdi.v6i1.12459.

[14]. Asgaryansyah, K., Elektro, J. T., & Mataram, U. (2024). Implementation of Linear Regression Algorithm to Measure the Level of Co2 Expenditure in Motor Vehicles. 2(3).