Analyzing Carbon Intensity Drivers and Forecasting Future Trajectories

A R Tahseen Jahan

I. Problem Motivation

Carbon intensity, measured as CO_2-equivalent emissions per unit of GDP, has become a key metric for assessing global progress toward decarbonization. As international climate commitments tighten, understanding the structural determinants of carbon intensity is increasingly important for policymakers, investors, and development institutions. Much of the existing empirical literature relies on traditional econometric models that assume linear and homogeneous relationships across countries. In reality, carbon intensity is shaped by complex and potentially nonlinear interactions among energy systems, levels of economic development, industrial composition, technological innovation, and natural-resource endowments. These relationships often differ substantially across regions and income groups, making conventional approaches limited in their ability to capture heterogeneity. Meanwhile, several decades of cross-country data from the World Bank create an opportunity to analyze how national characteristics drive carbon intensity dynamics over time. Machine-learning methods can complement standard econometric tools by revealing predictive patterns and interactions that may be overlooked in linear frameworks. Motivated by these opportunities and gaps, our project aims to: **First, we identify the key factors associated with carbon intensity across countries**, using variable-selection and tree-based ML methods; **second, we predict future levels of carbon intensity**, using a cleaned and harmonized global country-year panel.

II. Description of Data

1. Data Source

We use publicly available data from the **World Development Indicators (WDI)** provided by

the World Bank. The dataset covers **2000-2023**, depending on data availability.

2. Panel Construction

We constructed a unified country-year panel dataset, where each row corresponds to one country in one year. The response variable and all predictors were extracted from separate WDI indicator files. This produced a multi-dimensional panel with over 3,000 country-year observations.

3. Response Variable

Indicator name: **Carbon intensity of GDP** (kg CO_2-e per constant 2021 PPP dollar of GDP).

This variable is already standardized by GDP and comparable across countries.

4. Predictor Variables

Our predictor set spans six major structural factors discussed in the carbon-intensity literature.

5. Train-Test Split

Since the data is time-dependent, we avoid random sampling but using a **chronological split**:

- **Training set:** country-year observations from 2000-2022
- **Test set:** country-year observations in 2023 only

All models are trained on 2000-2022 and tested on 2023 in the current run. We plan to iteratively reduce the number of training years and evaluate multiple chronological splits to identify the optimal cutoff year. The final split will be chosen based on RMSE, MAE, and R^2.

III. Data Cleaning or Feature Engineering

To prepare the WDI dataset for analysis, we created a harmonized country-year panel combining all selected indicators from 2000-2023. Since the raw files differ in format and completeness, we took several steps to preprocess data and ensure consistency across countries and years, and we did not create any new variables in this process.

1. Reshaping and Merging Data

Each indicator was originally provided in a wide format (one row per country with multiple year columns). We reshaped all datasets into a long country-year format and merged them using country code and year.

2. Removing Invalid Zeros

Some indicators (R&D expenditure, natural resource rents, renewable electricity share) incorrectly record missing values as zeros. For variables where zero is not a meaningful economic value, these entries were treated as missing.

3. Handling Missing Values

Missing data varies across indicators and countries. To address this issue, we implemented a three-layer approach, outlined below: **First, time-series interpolation within each country.** For a given country and variable, if the first and last non-missing values were available, missing observations between them were filled using linear interpolation over time; second, **country-level mean imputation.** If an entire series for a given country was missing or almost entirely missing, interpolation was not possible. In such cases, we replaced missing values for that country-variable pair with the country's mean (when at least a few years were available) or, if necessary, with the overall sample mean in the next step**; third, global mean imputation.** For the small number of remaining missing values, we filled them with the global mean of the corresponding variable across all countries and years. This ensures that the final panel has no missing entries in the predictors and can be directly used by all ML algorithms.

IV. Machine Learning Methods Discussion

To identify the most effective predictive model for carbon intensity, we evaluated four classes of models, OLS, LASSO (with two regularization levels), Random Forest, and Gradient Boosting, using train/test splits and standard predictive performance metrics (RMSE, MAE, and R^2). The

goal of this section is to highlight how different model classes uncover complementary aspects of the data-generating process, and to compare their suitability for both inference and forecasting. Across all models tested, nonlinear tree-ensemble methods substantially outperform linear baselines because they can flexibly capture interactions among energy systems, economic structure, industrial value added, demographic profiles, innovation capacity, and natural-resource rents. Using 2023 as the forecasting anchor year, we select the Random Forest model as the preferred predictive learner for this iteration, while retaining OLS, LASSO, and Gradient Boosting as interpretable benchmarks to provide directional insight and robustness checks.

V. Justification of Final Model

To select a final forecasting model, we summarize performance across all methods in a comparison table, which reports train and test RMSE, MAE, and R^2 for each. Table 2 summarizes the results.

- OLS (Test RMSE = 0.372, Test $R^2 = 0.051$) and LASSO (Test $R^2 \approx 0.052$) provide similar out-of-sample performance, with LASSO offering further insights through variable selection and coefficient shrinkage.
- Gradient Boosting improves over linear models (Test RMSE = 0.094, Test $R^2 = 0.939$**)** but can be somewhat sensitive to tuning.
- Random Forest has the lowest test RMSE and MAE and the highest test R^2 on the 2023 data (Test $R^2 = 0.95$), which indicates the strongest out-of-sample predictive accuracy.

Given these findings, we select **Random Forest** as the final model. It offers:

1) Superior predictive performance on test data;
2) Flexibility to capture nonlinearities and interactions among key variables;
3) Built-in measures of variable importance that align well with prior domain knowledge

(energy use per capita and fossil fuel share typically emerge as highly important). Linear models and Gradient Boosting remain in valuable benchmarks and are used primarily for interpretability and robustness checks, while Random Forest serves mainly for prediction.

Table 1 Model Performance Summary

Model	Train RMSE	Train MAE	Train R^2	Test RMSE	Test MAE	Test R^2
OLS	0.40	0.13	0.10	0.37	0.13	0.05
LASSO (lambda.min)	0.40	0.13	0.10	0.37	0.13	0.05
LASSO (lambda.1se)	0.43	0.14	0.00	0.38	0.13	-0.01
Random Forest	0.15	0.04	0.88	0.09	0.05	0.95
Gradient Boosting	0.17	0.07	0.84	0.09	0.07	0.94

VI. Presentation of Results

Using the Random Forest model, we obtain a Test RMSE of 0.09 and a Test MAE of 0.05. Given that the actual 2023 carbon intensity of GDP has a mean of 0.19 and a standard deviation of 0.38, the model demonstrates strong predictive performance. This indicates that training the model on data from 2000 to 2022 provides a reliable basis for predicting carbon intensity in 2023.

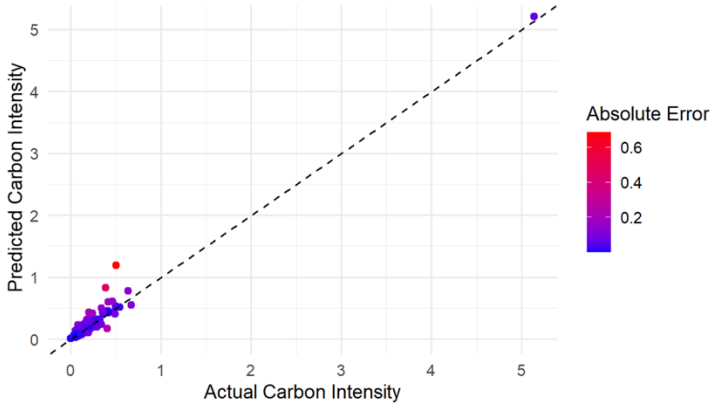


Figure 1 2023 Predicted and Actual Carbon Intensity

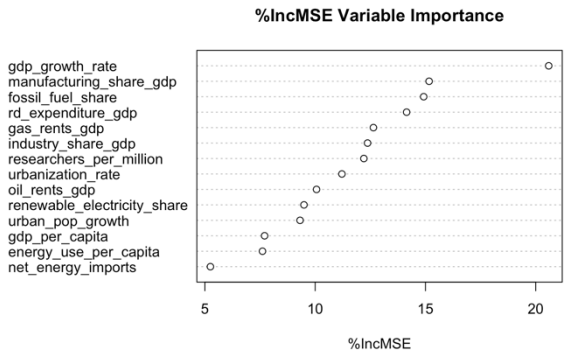


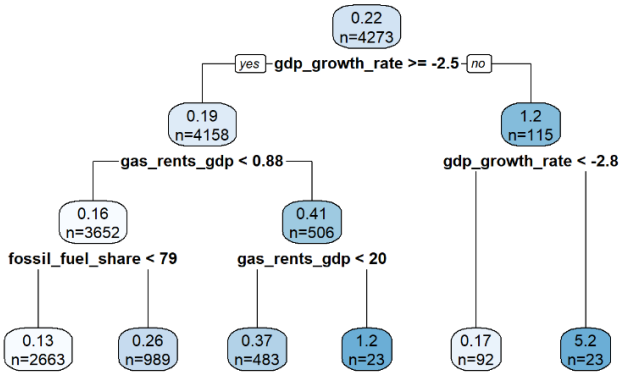
Figure 2 Variable Importance (Random Forest, %IncMSE)

In terms of variable importance measured by %IncMSE in the Random Forest model, the top five predictors are GDP growth rate (20.59), manufacturing share of GDP (15.17), fossil fuel share (14.92), R&D expenditure (14.15), and gas rents (14.31).

VII. Explanation of Final Model

The Random Forest constructs 500 decision trees, and the overall prediction is obtained by averaging the outcomes across all trees. Using a large ensemble of trees enhances prediction of stability and reduces variance. At each decision split within a tree, the model evaluates only a randomly selected subset of predictors rather than the full set. This design introduces randomness that lowers the correlation among individual trees and improves the model's generalization performance.

Figure 3 Example of the First Decision Tree



Also, the model is configured to compute variable importance scores, which assess the contribution of each predictor to reducing prediction error. These important measures allow the study to identify the most influential variables role in shaping carbon intensity outcomes.

VIII. Discussion of Results

1. Key Findings

The results demonstrate that the Random Forest model provides strong predictive performance in

estimating the carbon intensity of GDP. Compared with OLS, LASSO, and Gradient Boosting, the Random Forest achieves the lowest prediction errors on the test dataset, indicating that the model captures nonlinear relationships and complex interactions among variables more effectively than linear approaches. The variable-importance results, measured by the percentage increase in mean squared error (%IncMSE), further illuminate the key drivers of carbon intensity. %IncMSE captures how much the model prediction error rises when the information from a given variable is permuted or removed; higher values therefore indicate greater predictive influence. The finding that GDP growth, fossil-fuel share, and the manufacturing and industry components of GDP yield the highest %IncMSE scores suggests the model depends most heavily on economic and energy-related factors when predicting carbon intensity.

2. Bias and Fairness Considerations

Data Completeness and Imputation Bias: High-income, data-rich countries tend to report more complete and consistent indicators, whereas low-income or fragile states often have sparse or irregular data coverage that requires imputation. These imputation layers may disproportionately smooth sparse series, potentially leading to systematic under- or over-estimation of carbon intensity for data-poor countries.

Model-Related Biases in Random Forests: Tree models such as Random Forests inherently prioritize variables with higher variance and stronger predictive signal. This characteristic may reduce predictive accuracy for countries with atypical structures, such as small island states, conflict-affected economies, or extremely resource-dependent regions, since the model tends to learn dominant global patterns that may not fully represent such unique cases.

Looking Forward: To evaluate fairness across groups, future work should include subgroup error analysis (e.g., by income group, region) to identify systematic disparities. Incorporating

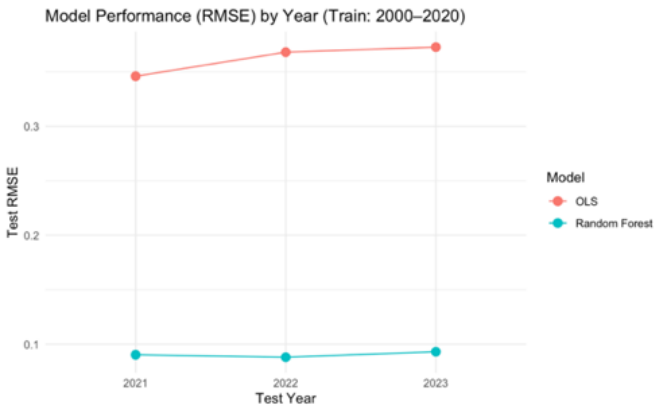
confidence intervals for country-level predictions can further mitigate misinterpretation.

3. Out-of-Sample Performance Discussion

Short-Term Out-of-Sample Performance: To assess near-term forecasting accuracy, we trained the model on 2000–2020 data and evaluated its predictions for 2021, 2022, and 2023. This design approximates how the model would perform in real-time forecasting. The RMSE results show that Random Forest consistently outperforms OLS and maintains stable accuracy across all three years. This pattern indicates that the Random Forests model effectively captures nonlinear effects and performs well in short-term out-of-sample performance.

Long-Term Out-of-Sample Performance: Although Random Forest performs well over short horizons, its accuracy depends on the stability of underlying economic and energy relationships. The model predicts most reliably within the historical range of observed data. Large structural changes, such as energy price shocks, geopolitical conflicts, rapid policy transitions, or shifts in industrial composition, may cause future carbon-intensity outcomes to diverge from historical patterns. As a result, predictive accuracy may. Our empirical results illustrate this pattern: the test RMSE of the Random Forest model increases gradually from 2021 to 2023.

Figure 4 Model Performance (RMSE) by Year (Train: 2000-2020)



Strategies to Improve Future OOS Robustness: To address declining performance in medium- and long-term predictions, several strategies can be implemented, including expanding the

training data with updated economic, energy, and emissions indicators, using rolling-window validation to better mimic real-time forecasting conditions, and combining Random Forest predictions with scenario-based modeling to account for uncertain future trajectories.

IX. Contribution to the Field

Advancing Methods for Cross-Country Carbon-Intensity Analysis: This study introduces machine-learning approaches, especially Random Forests, to carbon-intensity analysis, which could overcome the linearity and homogeneity assumptions of traditional econometric models. By capturing nonlinear interactions among economic structure, energy systems, and resource dependence, ML methods deliver substantially improved predictive accuracy. This advances methodological practice in global decarbonization research and demonstrates the value of flexible, data-driven models for understanding heterogeneous national emission pathways.

Building a Harmonized Global Dataset and Identifying Key Driver: The study develops a unified, reproducible country-year panel of more than 3,000 observations from 2000-2023, enabling robust cross-country comparisons. Using this dataset, the Random Forest model identifies GDP growth, fossil-fuel share, industrial value added, and resource rents as the most influential predictors of carbon intensity. These findings provide a clear, data-driven prioritization of structural factors shaping national emission efficiency and offer actionable insights for climate and development policy.

Delivering a Practical Framework for Forecasting Decarbonization Trends: The project produces a forecasting framework that can be updated annually to track global decarbonization progress. By highlighting which macro-structural variables most affect carbon intensity, the model offers practical guidance for policymakers, investors, and international organizations seeking to target high-impact reforms, such as reducing fossil-fuel dependence or modernizing

industrial systems, to accelerate low-carbon transitions.

X. Recommendations for Implementation

Use Forecasts to Guide Climate and Investment Planning: Policymakers can integrate annual carbon-intensity forecasts into national climate strategies, NDC tracking, and carbon budgeting, while investors can incorporate them into their portfolio risk analysis.

Prioritize Interventions in High-Impact Drivers: Since fossil fuel share, GDP growth, industrial structure, and resource rents strongly influence carbon intensity, governments and investors should target these areas, through renewable-energy expansion, industrial decarbonization, and support for cleaner production and energy efficiency.

Apply Sector-Specific Insights: Model diagnostics can inform better actions, such as cleaner manufacturing programs, incentive structures for energy-efficient technologies, and strategic investment in low-carbon sectors where emissions intensity is highest.

Strengthen Data Infrastructure and Annual Updates: Effective use of the model requires reliable, regularly updated economic and energy indicators. Governments should enhance data systems, and investors can integrate the updated model outputs into their analytical tools and risk-monitoring platforms.

Support Transition Finance and Public-Private Alignment: Both stakeholder groups can use results to coordinate transition financing, identifying priority sectors for green bonds, blended finance, and clean-technology investment aligned with countries' decarbonization pathways.

Incorporate Scenario Analysis for Strategic Decision-Making: Running model-based scenarios (changes in fossil fuel share or industrial composition) can help policymakers test policy packages and allow investors to evaluate climate-related risks and opportunities under alternative future conditions.

XI. Conclusion

This study applies machine-learning methods to analyze and forecast carbon intensity across more than 3,000 country-year observations, demonstrating that nonlinear ensemble models offer substantial improvements over traditional linear approaches. By leveraging Random Forests, we identify GDP growth, fossil-fuel dependence, manufacturing and industry value added, and natural-gas rents as the strongest structural determinants of carbon intensity. These findings reinforce the view that carbon-intensity outcomes arise from complex interactions among economic structures, energy systems, and resource endowments that cannot be fully captured through linear models. Our forecasting results suggest that machine-learning models can serve as valuable tools for short-term prediction and policy analysis, particularly when paired with robust data-cleaning pipelines and harmonized global indicators. At the same time, the study highlights several challenges, including data gaps, regional disparities, and potential structural shifts, that future research should address through improved data infrastructure, subgroup fairness analysis, and scenario-based modeling frameworks. Overall, this project contributes both a methodological advancement and a practical forecasting tool for policymakers, investors, and development partners seeking to monitor and accelerate global decarbonization.

Appendix

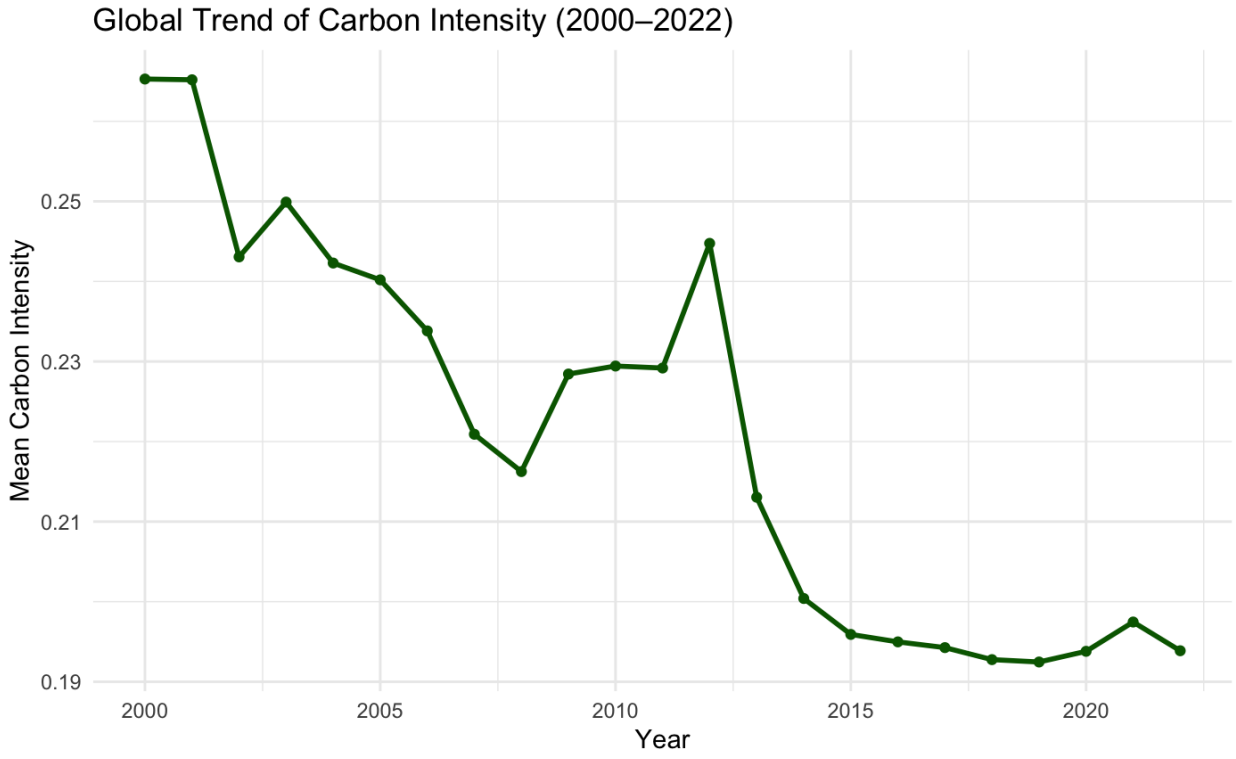
Appendix A Summary of Predictor Variables

Category	Variable Names	Variable Codes	Unit
Energy	Energy use	energy_use_per_capita	kg of oil equivalent per capita
	Energy imports	net_energy_imports	% use
	Fossil fuel share	fossil_fuel_share	% of total
	Electricity production from renewable sources, excluding hydroelectric	renewable_electricity_share	% of total
GDP	GDP per capita	gdp_per_capita	constant 2015 US$
	GDP growth	gdp_growth_rate	annual %
Industry	Industry (including construction), value added	industry_share_gdp	% of GDP
	Manufacturing, value added	manufacturing_share_gdp	% of GDP
Population	Urban population	urban_pop_growth	% of total population
	Urban population growth	urbanization_rate	annual %
Research	R&D expenditure	rd_expenditure_gdp	% of GDP
	Researchers per million	researchers_per_million	per million people
Energy rents	Oil rents	oil_rents_gdp	% of GDP
	Natural gas rents	gas_rents_gdp	% of GDP

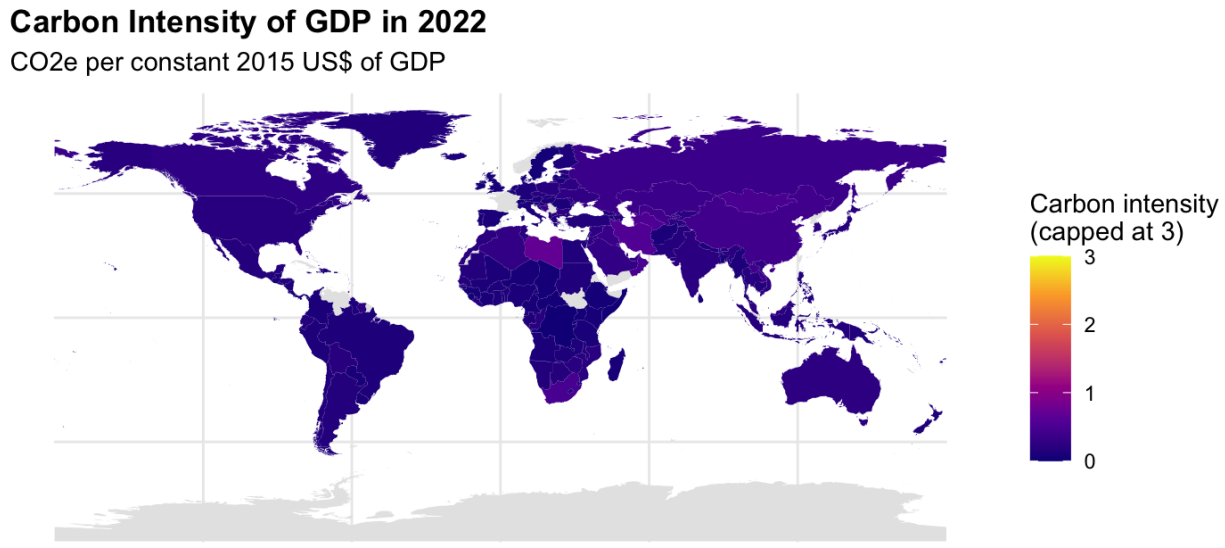
Appendix B Summary of Descriptive Statistics

variable	mean	sd	min	median	max
carbon_intensity	0.22	0.43	0.00	0.16	11.13
energy_use_per_capita	2208.19	2339.67	122.88	2105.18	16938.72
fossil_fuel_share	63.37	26.50	0.00	65.06	99.98
gas_rents_gdp	0.61	2.47	0.00	0.00	28.81
gdp_growth_rate	4.81	5.82	-14.28	4.34	58.08
gdp_per_capita	11678.43	17316.28	233.03	3885.30	109502.80
industry_share_gdp	27.04	11.52	4.25	25.99	84.80
manufacturing_share_gdp	13.43	6.82	0.91	13.42	44.98
net_energy_imports	-70.48	268.23	-2382.89	-49.65	218.73
oil_rents_gdp	5.10	12.57	0.00	0.01	82.78
rd_expenditure_gdp	0.96	0.54	0.04	0.98	3.83
renewable_electricity_share	2.21	9.88	-71.06	0.17	87.34
researchers_per_million	1557.29	679.29	14.76	1529.95	5102.41
urban_pop_growth	2.13	2.06	-9.89	2.02	8.88
urbanization_rate	53.94	23.61	8.25	53.42	100.00

Appendix C Global Trend of Carbon Intensity (2000-2022)



Appendix D Carbon Intensity of GDP in 2022



Appendix E Variables Correlation Heatmap