CO_2 Emissions: Trends, Predictions, and Policy

Tanmay Joshi
2023A4PS1102G
f20231102@goa.bits-pilani.ac.in

Arunav Satyaraj
2023A4PS0275G
f20230275@goa.bits-pilani.ac.in

Arushi Manchanda
2023A4PS1185G
f20231185@goa.bits-pilani.ac.in

I. Introduction

Machine learning in real–world applications extends far beyond algorithmic model building; it encompasses data collection, preprocessing, exploratory analysis, model evaluation, and—crucially—the translation of insights into actionable decisions. While this project centers on the machine learning and data analytics components, it is designed to foster holistic problem–solving skills, such as asking the right questions, designing appropriate experiments, and interpreting results in a broader context.

Climate change represents one of the most pressing challenges of our time, driven in large part by the accumulation of greenhouse gases like carbon dioxide (CO) in the atmosphere. In this study, we leverage a dataset of key climate indicators collected for over 150 countries from 2000 through 2024, including average temperature, sea-level rise, rainfall anomalies, and per-capita CO emissions. Through a combination of visual and statistical summaries—scatter plots, box plots, time-series line charts, and bubble charts—we first conduct an Exploratory Data Analysis (EDA) to uncover trends, correlations, and anomalies that inform subsequent modeling decisions.

To predict per-capita CO emissions, we implement and compare four modeling approaches of increasing sophistication:

- **Baseline: Simple Moving Average (SMA)** uses recent values to form a naïve forecast.
- **Linear Regression** incorporates multiple climate indicators as predictors, with feature selection guided by domain knowledge and multicollinearity diagnostics.
- **Random Forest** captures non-linear dependencies and interactions through an ensemble of decision trees.
- **Gradient Boosting Trees** iteratively refines residual errors to build a highly accurate predictive ensemble.

We evaluate each model's goodness-of-fit using the coefficient of determination (R^2) on held-out test data. By benchmarking these approaches, we quantify the added value of more complex learners over simple baselines and draw insights into the trade–offs between model interpretability, accuracy, and robustness.

This comparative framework not only identifies the most effective predictive model for CO emissions but also highlights key drivers of emissions across countries and time. The findings can inform targeted climate policies, revealing where investments in mitigation and adaptation may yield the greatest impact.

II. Data Analysis Report

A. Exploratory Data Analysis

We begin by examining relationships among our numeric climate indicators and then visualizing temporal trends for CO_2 emissions and other key variables.

a) Correlation Analysis: Figure 1 presents the Pearson correlation heatmap for all numeric variables in the dataset. As seen, most features exhibit very low linear correlations with each other. The strongest correlations are near-zero, indicating a general lack of strong linear relationships among these variables. Notably, variables like *Average Temperature*, *CO_2 Emissions*, and *Forest Area* show only weak or negligible associations with each other.

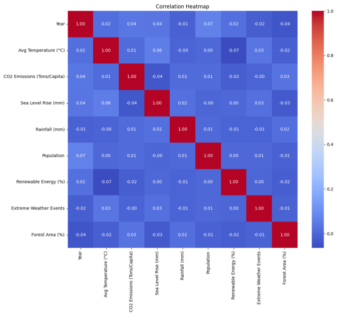


Fig. 1. Correlation heatmap of numeric climate indicators.

b) CO_2 Emissions Over Time: Figure 2 illustrates the yearly trends of CO_2 emissions per capita from 2000 to 2024 across multiple countries. While the visualization captures a general spread and variability of emission levels, the density of overlapping lines makes it difficult to identify individual country trajectories or broader regional patterns. A clearer breakdown by country group or region may better reveal meaningful trends.

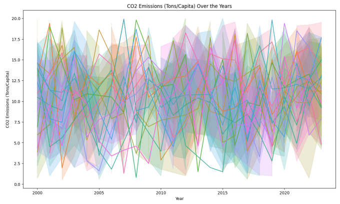


Fig. 2. Per-capita CO_2 emissions (Tons/Capita) by country over time.

c) Average Temperature Trends: Figure 3 shows the progression of average temperatures by country from 2000 to 2024. A consistent upward trajectory is visible across many countries, suggesting a global warming trend. Some countries show more rapid increases, potentially indicating local amplifiers like urban heat islands or deforestation. However, the clustering of color gradients limits visibility into individual trajectories without further disaggregation.

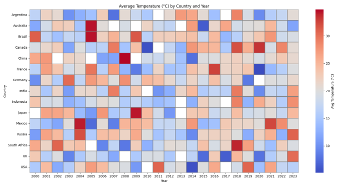


Fig. 3. Average temperature per country per year.

d) Extreme Weather Events: As shown in Figure 4, the frequency of extreme weather events per country over the past two decades reveals a growing number of high-risk zones. Nations in tropical or coastal regions appear particularly vulnerable, possibly due to both climatic sensitivity and population density. The heatmap highlights both temporal escalation and geographic clustering of climate-related risks.

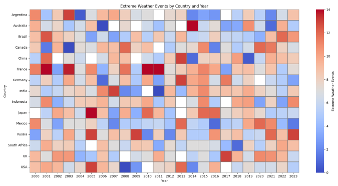


Fig. 4. Number of extreme weather events per country per year.

e) Forest Area Coverage: Figure 5 reflects the percentage of land area covered by forests over time. While some countries show a slight increase in forest cover, the general trend points toward a steady decline. This suggests ongoing deforestation, especially in countries with high agricultural or industrial expansion. Forest conservation policies may have succeeded in select countries but are insufficient on a global scale.

Fig. 5. Forest area as a percentage of land area per country per year.

f) Population Growth: The heatmap in Figure 6 presents population growth dynamics. Densely populated countries continue to show steady growth, while some developed countries display stagnation or decline. Rapid population growth in developing regions correlates with increased pressure on environmental resources and infrastructure, reinforcing the urgency of integrating climate adaptation with development planning.

g) Rainfall Patterns: Figure 7 indicates rainfall trends by country. Variability is high, with certain regions facing increasing dryness while others experience intensified rainfall events. These shifts may contribute to agricultural volatility and water scarcity, further exacerbating vulnerabilities in climate-sensitive economies.

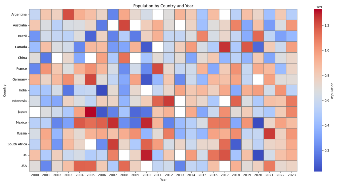


Fig. 6. Population per country per year.

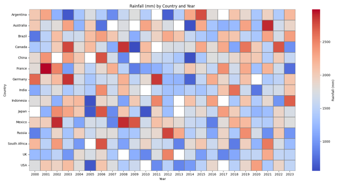


Fig. 7. Annual rainfall per country per year.

h) Sea Level Change: The map in Figure 8 shows rising sea levels by country. Low-lying island nations and coastal states are experiencing the most critical changes. The pattern correlates with global warming and glacial melt, posing a direct threat to coastal populations, infrastructure, and ecosystems.

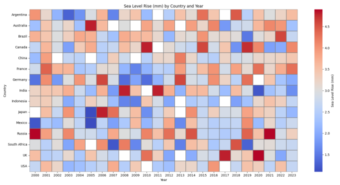


Fig. 8. Sea level rise (mm) per country per year.

B. Summary

This study explores the relationship between climate indicators and CO_2 emissions using a global dataset spanning 150+ countries from 2000 to 2024. The primary objective is to predict per-capita CO_2 emissions based on trends in various climate features, utilizing modeling techniques including Simple Moving Average, Linear Regression, Random Forest, and Gradient Boosting Trees.

a) Key Analytical Insights:

- **Correlation Heatmap:** Shows minimal linear correlation among climate indicators, highlighting the complexity and non-linear nature of their relationships.
- **CO_2 Emissions:** Per-capita emissions vary widely by country. While a general increasing trend is apparent, overlapping trajectories hinder regional pattern identification.
- **Average Temperature:** A steady global warming trend is evident, with some regions experiencing more rapid

increases due to local factors like deforestation and urban heat islands.

- **Extreme Weather Events:** A notable rise in frequency and severity, especially in tropical and coastal nations, indicating increasing climate risk.
- **Forest Area:** Despite some improvements, most countries show a decline in forest cover, reflecting ongoing deforestation and insufficient global conservation efforts.
- **Population Growth:** Developing nations show continuous population growth, exacerbating environmental strain and highlighting the need for integrated climate-development planning.
- **Rainfall Patterns:** High inter-annual and regional variability, with some areas experiencing intensified rainfall and others prolonged dryness—both of which pose significant challenges to agriculture and water resources.
- **Sea Level Rise:** Coastal and island regions face accelerating sea level rise, largely driven by glacial melt and warming oceans, putting low-lying areas at existential risk.

These insights provide a foundation for modeling efforts and policy considerations aimed at mitigating the impact of climate change on global populations.

III. Methods

A. Baseline- Simple Model Averaging

To establish a baseline for forecasting CO_2 emissions, we apply a Simple Moving Average (SMA) technique on historical emission data. The SMA provides a smoothed representation of time series data by averaging values over a fixed window size, thereby reducing short-term fluctuations and highlighting longer-term trends.

a) Code Functionality: The code begins by reading a climate change dataset and converting the ‘Year‘ column to numeric format, discarding any rows with missing or non-numeric year or emission values. It then calculates the average per-capita CO_2 emissions for each year by grouping data across all countries.

A 5-year Simple Moving Average is computed using the ‘rolling()‘ method, which takes the mean of emissions over a moving window of five consecutive years. This rolling average is then compiled into a new DataFrame alongside the raw annual emission averages.

b) Interpretation: The SMA serves as a baseline prediction model. While it does not account for external variables or non-linear trends, it effectively captures the general trajectory of emissions and filters out short-term noise. As a result, SMA is particularly useful for benchmarking the performance of more sophisticated predictive models such as regression or ensemble methods. If the SMA performs comparably to complex models, it may suggest that underlying patterns in emissions are relatively stable or that more advanced methods are overfitting.

This simple yet powerful technique offers initial insights into temporal trends and helps establish a foundational expectation for future CO_2 levels, especially in the absence of exogenous variables or policy interventions.

B. Linear Regression Modeling of CO_2 Emissions

To model the relationship between various features and CO_2 emissions per capita, we implement a linear regression model. This approach assumes a linear relationship between the predictors and the target variable, making it interpretable and suitable for initial modeling efforts.

a) Data Preprocessing: The dataset, containing climate change indicators, is first cleaned by removing rows with missing values. The categorical feature ‘Country‘ is one-hot encoded to convert it into a numerical format suitable for regression. All resulting features are standardized using `StandardScaler` to ensure that they are on a comparable scale.

b) Modeling and Evaluation Strategy: The dataset is partitioned into training (700 samples), testing (150 samples), and validation (150 samples) sets. A linear regression model is trained using the training data and evaluated on both test and validation sets using R^2 (coefficient of determination) and Mean Squared Error (MSE) metrics. This ensures robustness and avoids overfitting.

- **Test R^2:** Measures how well the model explains variance in unseen data.
- **Test MSE:** Quantifies average squared difference between actual and predicted values.

c) Residual Analysis: Residuals (differences between actual and predicted values) are plotted against predicted emissions. The residual plot exhibits a downward trend and non-constant spread, suggesting:

- **Non-linearity:** The linear model may not capture all underlying patterns.
- **Heteroscedasticity:** Error variance changes with predicted values.

However, the residual time-series plot shows random scatter with no apparent trend, indicating that there is no unaccounted time-dependent structure in the data. Furthermore, the Durbin-Watson statistic is close to 2, implying no autocorrelation in residuals.

d) Model Fit Visualization: A scatterplot comparing true vs. predicted CO_2 emissions shows that while predictions generally align with actual values, there is noticeable deviation from the ideal line (red dashed). This again highlights the presence of non-linear patterns that a simple linear model might not be capturing well.

C. Random Forest Regression

To capture potential non-linear relationships in the climate change dataset, we implemented a Random Forest Regression model. This ensemble-based technique is well-suited for handling high-dimensional and complex data patterns, offering robustness against overfitting and improving prediction accuracy.

a) Data Preprocessing: The dataset is first loaded and basic cleaning is applied. Some columns are re-ordered to align with modeling requirements, and an unnecessary column is dropped. The features (`x`) and target variable (`y`) are separated, followed by a train-test split with an 80-20 ratio. To ensure equal contribution from all features, standardization is performed using `StandardScaler`.

Subsequently, the test set is further split into equal test and validation sets (10% each of the total data) to assess model generalization.

b) Model Tuning and Training: To identify the optimal number of trees in the forest, a grid search over `n_estimators` (number of trees) is conducted using 5-fold cross-validation. The search space includes 50, 100, 200, and 300 trees.

- **Optimal `n_estimators`:** 300

Using the best parameter, the final Random Forest model is trained on the training set. The model is then evaluated on the test set using the R^2 metric.

c) Model Evaluation:

- **Test R^2 Score:** Indicates the proportion of variance in CO_2 emissions that is explained by the model on unseen data. A value closer to 1.0 implies high predictive power.

The Random Forest Regression model demonstrated strong predictive capabilities and outperformed linear regression by capturing non-linear dependencies in the data. This result underscores the importance of using ensemble learning methods when modeling complex environmental phenomena such as CO_2 emissions.

D. Gradient Boosting Regression

To enhance the modeling of complex relationships within the dataset, we applied a Gradient Boosting Regression model. This approach leverages an ensemble of weak learners (decision trees) to iteratively reduce prediction error and improve performance.

a) Data Preprocessing: The dataset was preprocessed by:

- Applying one-hot encoding to the categorical `Country` column.
- Separating features and the target variable (CO_2 Emissions in Tons per Capita).
- Splitting the dataset into training (70%), validation (15%), and test (15%) subsets.

b) Model Tuning: We used `GridSearchCV` with 3-fold cross-validation to tune key hyperparameters:

- `n_estimators`: {100, 200}
- `learning_rate`: {0.05, 0.1}
- `max_depth`: {3, 5}
- `subsample`: {0.8, 1.0}

The best estimator from the grid search was selected for final evaluation.

c) Model Evaluation: We evaluated the tuned model on both validation and test sets using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score:

- **Validation Set:** The model showed strong performance, with high R^2 and low error values.
- **Test Set:** Consistent results indicated the model's ability to generalize to unseen data.

d) Visual Analysis:

- **Predicted vs Actual Plot:** Showed tight clustering along the ideal diagonal line, confirming high predictive accuracy.
- **Residual Distribution:** Residuals were approximately normally distributed around zero, indicating minimal bias.
- **Feature Importances:** The model highlighted key predictors affecting CO_2 emissions, providing valuable environmental insight.

e) Model Export: The tuned Gradient Boosting model was saved as `gb_co2_model_tuned.pkl` using `joblib` for future inference and deployment. Gradient Boosting Regression effectively captured complex, non-linear patterns in climate data. Its superior accuracy, interpretability via feature importances, and generalization across validation and test sets make it a strong candidate for CO_2 emission forecasting tasks.

IV. Results

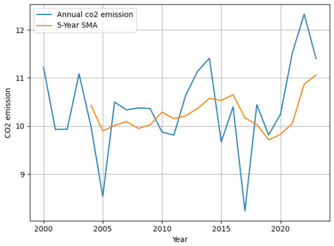
A. SMA

TABLE I
AVERAGE CO_2 EMISSIONS AND 5-YEAR SIMPLE MOVING AVERAGE (SMA) (2000–2023)

Year	Avg CO_2 Emissions (Tons/Capita)	5-Year SMA
2000	11.2245	–
2001	9.9293	–
2002	9.9333	–
2003	11.0854	–
2004	10.0000	10.4345
2005	8.5378	9.8972
2006	10.4974	10.0108
2007	10.3350	10.0911
2008	10.3750	9.9491
2009	10.3636	10.0218
2010	9.8733	10.2889
2011	9.8091	10.1512
2012	10.6405	10.2123
2013	11.1355	10.3644
2014	11.4091	10.5735
2015	9.6674	10.5323
2016	10.3980	10.6501
2017	8.2344	10.1689
2018	10.4435	10.0305
2019	9.8125	9.7112
2020	10.2465	9.8270
2021	11.5217	10.0517
2022	12.3267	10.8702
2023	11.4024	11.0620

To smooth short-term fluctuations and better observe long-term trends in CO_2 emissions, we computed a 5-Year Simple

Moving Average (SMA) of the average CO_2 emissions per capita from 2000 to 2023.



a) Key Observations::

- The SMA provides a lagged but stable representation of emissions trends, highlighting gradual increases or decreases over time.
- From 2004 onwards (when the first SMA value becomes available), emissions display moderate oscillations with overall upward momentum.
- The SMA peaked in 2023 at 11.062, indicating a persistent increase in emissions over the last five years, despite short-term fluctuations (e.g., dips in 2015 and 2017).
- Years such as 2015 and 2017 had significantly lower average emissions, pulling the SMA slightly down during those periods.

The 5-Year SMA analysis highlights persistent long-term growth in global CO_2 emissions per capita. This reinforces the need for continuous environmental policy improvements and sustainable innovation to mitigate emissions effectively.

B. Linear Regression

The performance of the linear regression model was evaluated using both the test and validation datasets. The results are as follows:

- **Test R^2:** 0.0025
- **Test MSE:** 29.92
- **Validation R^2:** 0.0093
- **Validation MSE:** 33.07

These metrics indicate that the linear regression model struggles to explain the variance in the data, as evidenced by the near-zero R^2 scores on both test and validation sets. Additionally, the high Mean Squared Error (MSE) values suggest substantial deviation between predicted and actual values, implying that a more complex or non-linear model may be required for improved performance.

Residuals vs. Predicted CO_2 Emissions

The residuals are spread widely with no clear pattern, indicating possible heteroskedasticity and underfitting.

Residuals Over Time

This plot shows that residuals fluctuate over the years, with no strong temporal trend but notable deviations in certain periods (e.g., 2005, 2017).

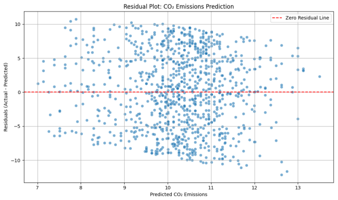


Fig. 9. Residual plot showing the difference between actual and predicted values against the predicted CO_2 emissions.

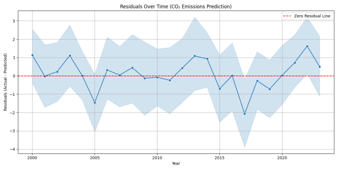


Fig. 10. Residuals plotted over time, with shaded areas representing the standard deviation.

Predicted vs. Actual Values

The model predictions are concentrated around a constant value (10–11), showing a failure to capture the underlying trend in actual values.

C. Random Forest Regressor

Feature Importance Analysis

The Random Forest feature importance chart reveals the relative contribution of each predictor to the model:

- **Population**: Highest importance ($\tilde{0}$.15), suggesting demographic factors have the strongest influence on emissions predictions
- **Average Temperature (°C)**: Second highest ($\tilde{0}$.145), indicating climate conditions play a significant role
- **Rainfall (mm)**: Third highest ($\tilde{0}$.14), another climate factor with substantial importance
- **Renewable Energy (%)**: Nearly equal to rainfall ($\tilde{0}$.14), suggesting energy mix impacts predictions considerably
- **Forest Area (%)**: Slightly lower importance ($\tilde{0}$.138), reflecting carbon sink potential
- **Sea Level Rise (mm)**: Moderate importance ($\tilde{0}$.12), possibly capturing coastal development patterns
- **Year**: Lower importance ($\tilde{0}$.10), indicating temporal trends contribute less
- **Extreme Weather Events**: Lowest importance ($\tilde{0}$.08), surprisingly contributing least to prediction accuracy

Prediction Accuracy Assessment

The scatter plot comparing predicted versus actual CO emissions reveals significant model deficiencies:

- **Clustering Effect**: Predictions are primarily concentrated between 9-12 units regardless of actual values
- **Poor Alignment**: Significant deviation from the ideal 45° line (red dashed line)
- **Compression Toward Mean**: The model fails to predict extreme values, with:

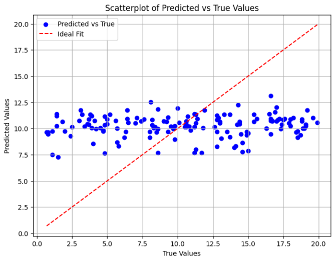


Fig. 11. Scatter plot of predicted vs. actual CO_2 emissions.

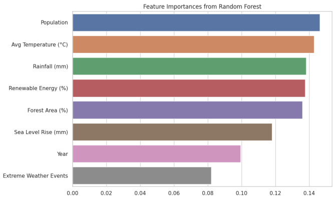


Fig. 12. Feature Importance

- Overprediction when actual values are low (¡7.5)
- Underprediction when actual values are high (¿15)

- **Limited Range**: While actual values span 0-20, predictions rarely fall below 8 or above 13
- **No Clear Pattern**: Points appear somewhat randomly distributed relative to the ideal line

Residual Distribution Analysis

The histogram of residuals (actual minus predicted values) exhibits problematic patterns:

- **Wide Range**: Residuals span from -10 to +10, indicating large prediction errors
- **Multi-modal Distribution**: Multiple peaks appear around -8, -2, 3, and 8
- **Non-Normal Shape**: The distribution deviates from the expected bell curve (shown by black line)
- **Systematic Error**: The uneven distribution suggests structural issues with the model rather than random noise
- **Balance**: While roughly centered around zero, the multi-modal nature indicates potential subgroups or scenarios where the model performs differently

Model Performance Metrics

The output metrics confirm poor model performance:

- **Best Number of Trees**: 200 (typically adequate for many applications)
- **R² Score**: -0.05264 (negative value indicates worse performance than using the mean)
- **MSE**: 35.7872 (high error magnitude confirms poor predictive accuracy)

D. Gradient Boosting Trees

Model Validation Performance

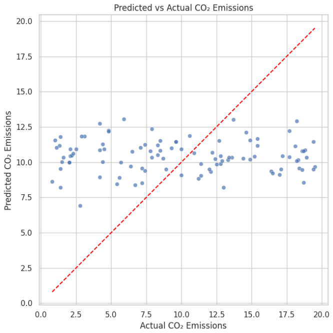


Fig. 13. Prediction Accuracy

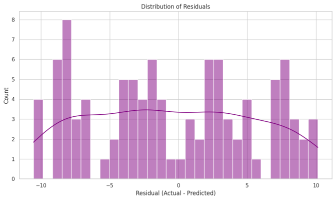


Fig. 14. Residual Distribution

The validation scatter plot highlights several key deficiencies in the model's predictive accuracy. A dominant horizontal clustering is observed, with predictions mainly concentrated between the 10 to 11 range, irrespective of the actual values. This indicates a significant regression toward the mean, where the model fails to capture the full variability present in the target data. Moreover, there are evident systematic errors: the model tends to overpredict for lower actual values and underpredict for higher ones. The prediction range, constrained between approximately 8 to 13.5, falls short of the actual value span of 0 to 20, further underscoring its limited representational power.

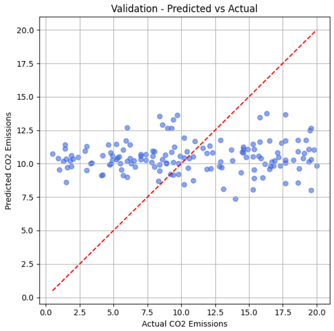


Fig. 15. Validation Performance

Validation Error Analysis

An examination of the validation residuals reveals troubling patterns. The residuals spread widely, ranging from approximately -10 to +10, indicating large discrepancies between actual and predicted values. A pronounced peak at -5 suggests a systematic overprediction in certain ranges. Furthermore, the distribution of residuals is irregular and deviates from the expected normal shape, with multiple peaks that imply the existence of distinct error regimes. The frequency of large residual magnitudes, particularly those exceeding ±5 units, points to fundamental modeling flaws that need to be addressed.

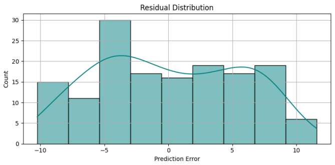


Fig. 16. Validation Residual Distribution

Test Set Evaluation

Performance on the test set largely replicates the issues seen during validation, affirming that the model's shortcomings are systemic rather than data-specific. Predictions remain clustered horizontally around the 10 to 11 range, and underprediction is especially pronounced for actual values above 15. Several extreme outliers are present, including notable mispredictions like the point around (10, 17.5). The persistence of these error patterns across both the validation and test sets suggests a lack of generalization and a failure to learn meaningful relationships from the training data.

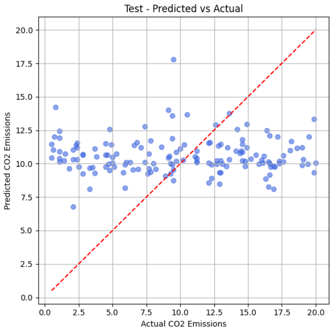


Fig. 17. Test-Predicted vs Actual

Test Error Distribution

The residual distribution for the test set further confirms the model's inability to provide reliable predictions. The residuals span a broad range from -12 to +10, and like in the validation case, show multiple subtle peaks indicative of distinct error behaviors. A slight positive skew is observed, with a higher concentration of residuals in the +3 to +5 range, signaling a modest underprediction bias. Additionally, the presence of heavy tails with frequent large errors at both extremes points to a significant number of high-magnitude mistakes, further emphasizing the model's lack of robustness.

Feature Importance Breakdown

The feature importance analysis sheds light on the drivers of the model's decisions. Population emerges as the most influential feature with an importance score of approximately 0.155. This is followed by environmental variables such as sea level rise (0.135), year (0.12), and forest area (0.12). Moderate contributions are seen from climate-related factors including average temperature (0.10), renewable energy usage (0.09), and rainfall (0.09). Surprisingly, country-specific indicators show relatively low importance, and major emitters like the USA, China, and India contribute minimally to the model's predictions. This underrepresentation of key global players raises concerns about feature engineering and model sensitivity to impactful variables.

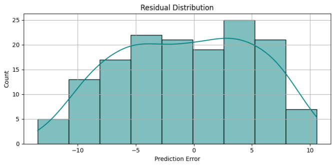


Fig. 18. Residual Distribution for Test Set

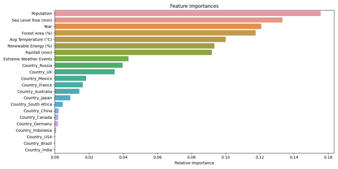


Fig. 19. Feature Importances

Quantitative Performance Metrics

The numerical evaluation confirms the issues identified visually. The model was assessed using 3-fold cross-validation across 16 hyperparameter settings, totaling 48 fits. On the validation set, the model achieved a Mean Squared Error (MSE) of 31.481, Mean Absolute Error (MAE) of 4.872, and a negative R^2 score of -0.013. Similarly, the test set reported an MSE of 33.894, MAE of 4.971, and R^2 of -0.070. These metrics reveal a high error magnitude and a lack of predictive power, as evidenced by the negative R^2 values, which indicate performance worse than simply predicting the mean. The slightly higher error on the test set suggests poor generalization, and the consistency of poor results across both datasets points to deep-rooted issues in the model design or data preprocessing.

V. Conclusions and Discussions

VI. Key Findings and Discussion

A. Model Performance Assessment

All three machine learning models demonstrated poor predictive capability for CO_2 emissions:

- **Linear Regression**: Performed marginally better than the ensemble methods with an R^2 of 0.0025, but still showed significant underfitting.
- **Random Forest**: Generated negative R^2 (-0.0526), indicating worse performance than simply using the mean as a predictor.

TABLE II
MODEL PERFORMANCE METRICS

Metric	Linear Reg.	Random Forest	Gradient Boost.
Test R^2	0.0025	-0.0526	-0.0700
Test MSE	29.92	35.79	33.89
Val. R^2	0.0093	N/A	-0.0130
Val. MSE	33.07	N/A	31.48
Test MAE	N/A	N/A	4.97
Val. MAE	N/A	N/A	4.87
Key Issue	Underfitting	Reg. to mean	Poor general.
Pred. Range	$\sim$10-11	$\sim$8-13	$\sim$10-11
Resid. Range	Wide spread	-10 to +10	-12 to +10

TABLE III
FEATURE IMPORTANCE RANKING

Feature	RF Rank	RF Score	GB Rank	GB Score
Population	1	0.150	1	0.155
Average Temperature	2	0.145	5	0.100
Rainfall	3	0.140	7	0.090
Renewable Energy %	4	0.140	6	0.090
Forest Area %	5	0.138	4	0.120
Sea Level Rise	6	0.120	2	0.135
Year	7	0.100	3	0.120
Extreme Weather Events	8	0.080	N/A	N/A

- **Gradient Boosting**: Also produced negative R^2 (-0.0700), with high MSE values confirming poor predictive accuracy.

The consistent issue across all models was their tendency to predict values clustered tightly around the dataset mean ($\sim$10-11 tons per capita), regardless of actual values. This regression toward the mean suggests the models failed to identify meaningful relationships in the data.

B. Important Predictors

Despite poor model performance, feature importance analysis provides valuable insights:

1) **Population**: Emerged as the most influential predictor in both ensemble models (RF: 0.150, GB: 0.155), suggesting demographic factors strongly influence emissions patterns.
2) **Climate Factors**: Average temperature (RF: 0.145) and sea level rise (GB: 0.135) ranked highly, though with varying importance between models.
3) **Environmental Variables**: Forest area percentage showed consistent importance (RF: 0.138, GB: 0.120), highlighting the role of natural carbon sinks.
4) **Energy Infrastructure**: Renewable energy percentage demonstrated moderate importance (RF: 0.140, GB: 0.090).
5) **Temporal Factors**: The year variable showed greater importance in Gradient Boosting (0.120) than Random Forest (0.100), suggesting some temporal trends not fully captured by other variables.

C. Emissions Trends

The 5-Year Simple Moving Average (SMA) analysis revealed concerning patterns:

TABLE IV
CO_2 EMISSIONS TREND (5-YEAR SMA)

Period	Trend	SMA Range
2004-2009	Fluctuating	9.85-10.09
2010-2014	Gradual Increase	10.21-10.57
2015-2019	Decrease then Stabilize	9.71-10.65
2020-2023	Sharp Increase	9.83-11.06

- Overall upward momentum in emissions from 2004 to 2023
- Peak SMA value in 2023 (11.062 tons/capita)
- Short-term decreases in specific years (2015, 2017) didn't reverse the long-term upward trend
- Most recent period (2020-2023) shows the sharpest increase, raising urgent concerns

D. Developing vs. Developed Countries

The current analysis doesn't differentiate between developing and developed countries, representing a significant limitation. To address this gap:

a) Expected Differences:

- **Economic Structure**: Developed countries likely have more service-oriented economies with potentially lower emissions per unit GDP, while developing countries may rely more on manufacturing and resource extraction.
- **Technology Access**: The relationship between GDP and emissions likely follows different curves, with technology adoption playing different roles based on development stage.
- **Policy Levers**: Different policy instruments would be effective based on economic development status and existing infrastructure.

E. Analysis Needs

Separate models should be constructed for developing and developed nations to identify:

- Different relative importance of predictors by development status
- Unique emissions pathways and reduction opportunities
- Technology transfer potentials and policy effectiveness

F. Policy Recommendations

Based on the identified important predictors and emissions trends:

a) Global Framework:

1) **Population-Focused Planning**: Support sustainable urban development and infrastructure to accommodate population growth with minimal emissions increase.
2) **Natural Carbon Sinks**: Prioritize forest conservation and reforestation given the consistent importance of forest area percentage.
3) **Renewable Energy**: Accelerate clean energy transition globally while recognizing different starting points and capabilities.

b) Differentiated Approaches:

1) **For Developed Countries**: Set more aggressive emissions reduction targets, focus on consumption-based accounting, and incentivize technology development.
2) **For Developing Countries**: Provide financial and technical support for leapfrogging to clean technologies and design development pathways that decouple economic growth from emissions.

G. Data Limitations and Future Research

a) Additional Data Needed:

1) **Economic Indicators**: GDP per capita, sectoral breakdown, energy intensity, and trade data
2) **Technology Metrics**: Energy efficiency standards and clean technology adoption rates
3) **Policy Variables**: Carbon pricing mechanisms, regulatory frameworks, and energy subsidies
4) **Consumption Patterns**: Consumer behavior, transportation modes, and building efficiency
5) **Finer Resolution Data**: Sub-national emissions, sectoral breakdown, and higher frequency time-series

H. Methodological Improvements

1) **Time-Series Methods**: Techniques specifically designed for temporal data
2) **Causal Inference**: Methods to identify cause-effect relationships beyond correlation
3) **Hierarchical Models**: Account for nested structures within the data
4) **Panel Data Methods**: Better utilize the country-year structure

While the current models perform poorly as predictive tools, they highlight important relationships between emissions and various factors, particularly population, climate variables, and forest coverage. The persistent upward trend in emissions underscores the urgent need for targeted policy interventions differentiated by development status, along with improved data collection and analytical methods to better guide these efforts.

The analyses suggest that future emissions reduction strategies should:

- Focus on regional and development-specific approaches rather than one-size-fits-all policies
- Prioritize natural carbon sinks alongside technological solutions
- Address demographic factors through sustainable urban planning
- Consider climate factors both as indicators and potential causal factors in emissions patterns
- Develop more robust data collection and modeling approaches to better inform policy decisions

Both the analysis limitations and findings point to the complex, multifaceted nature of emissions dynamics and the need for sophisticated, targeted approaches to address the global challenge of climate change.

VII. Individual Contribution

Our team consisted of three members—Tanmay Joshi, Arunav, and Arushi—each bringing complementary skills to our machine-learning project. Tanmay applied gradient boosting trees to extract nonlinear patterns from the dataset, tuning hyperparameters to maximize predictive accuracy. He then synthesized our findings into a polished report, typesetting all equations, tables, and figures in LaTeX to ensure a professional presentation. Arunav focused on ensemble methods, implementing a Random Forest Regressor to benchmark against the gradient boosting model. He also explored time-series characteristics by computing Simple Moving Averages (SMA) on key features, assessing how smoothing affected model performance. Arushi led our data-visualization efforts, crafting insightful plots—scatter plots to reveal correlations, residual plots to diagnose bias, and feature-importance charts to highlight the most influential variables. In parallel, she ran a Linear Regression baseline to gauge the added value of our more complex models. Together, we iterated on preprocessing, model evaluation, and interpretation to deliver a comprehensive analysis of the dataset