# Predictive Modeling of $CO_2$ emissions

Anand Shankar Hariharan
*2022B5A70941G*

S Mohammed Ashraf
*2022A7PS0645G*

***Abstract*—This project uses machine learning to predict $CO_2$ emissions per capita using various climate and economic factors. We examined data from 2000-2023 across multiple countries, looking at indicators like temperature, rainfall, renewable energy use, and forest coverage.**
**We implemented several models including Simple Moving Average as a baseline, Linear Regression, and more advanced approaches like Random Forest, XGBoost, and Support Vector Regression (SVR).**
**Interestingly, all models had negative $R^2$ values, showing how challenging it is to predict $CO_2$ emissions accurately. The SVR model performed best with the lowest error rate (MSE: 30.92). Our analysis of feature importance showed that renewable energy percentage, rainfall patterns, forest coverage, and temperature were the most significant predictors.**
**We also found performance differences when modeling developed vs developing countries, suggesting that different policy approaches might be needed for different economic contexts when working to reduce emissions.**

## I. Introduction

Climate change is one of the biggest challenges we face today. While scientists clearly understand that $CO_2$ emissions contribute to global warming, it's much harder to model exactly what factors drive emission patterns in different countries. The project uses ML to tackle this problem by analyzing and predicting $CO_2$ emissions per capita across various nations.

Being able to predict emissions accurately could help policymakers. Good models could show which policies would have the biggest impact, help governments use resources more effectively, and suggest different approaches for countries at different economic stages. But as our results will show, building reliable prediction models is tough because so many different factors are involved.

In this project, we aimed to:

1. Build and compare several different models for predicting $CO_2$ emissions
2. Figure out which factors have the strongest influence on emission levels
3. Look at differences between developed and developing countries
4. Suggest some data-backed policy recommendations that could help reduce emissions

## II. Data and Methodology

The dataset contains climate change indicators collected from multiple countries between 2000 and 2023. It includes 1,000 observations across 10 variables:

- Year (2000-2023)
- Country
- Average Temperature (°C)
- $CO_2$ Emissions (Tons/Capita) - target variable
- Sea Level Rise (mm)
- Rainfall (mm)
- Population
- Renewable Energy (%)
- Extreme Weather Events
- Forest Area (%)

Countries were classified as "Developed" or "Developing" based on common economic classifications.

### A. Exploratory Data Analysis

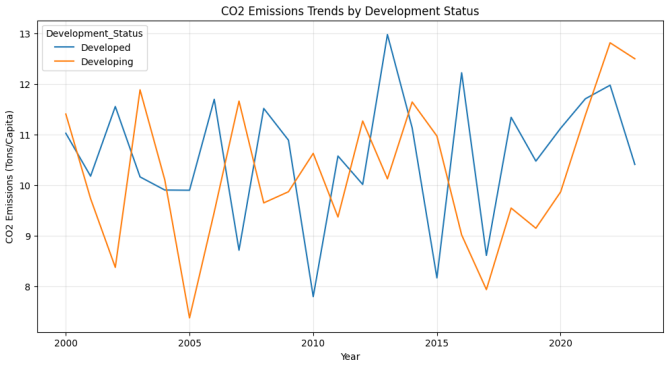


Fig. 1. CO2 emissions trends by development status.

Initial analysis revealed some patterns in the data:

- The dataset is balanced across years with measurements from 2000 to 2023.
- $CO_2$ emissions range from 0.5 to 20 tons per capita, with a mean of 10.43.
- Average temperature ranges from 5°C to 34.9°C.
- There is minimal correlation between most climate variables (see Fig 2).
- Temporal trends in $CO_2$ emissions show different patterns between developed and developing nations.

### B. Feature Engineering

To account for the time-series nature of the data, we implemented some preprocessing steps:

1) Time-lagged features: For each country, we created lagged variables of climate indicators (previous year's temperature, rainfall, sea level, etc.).

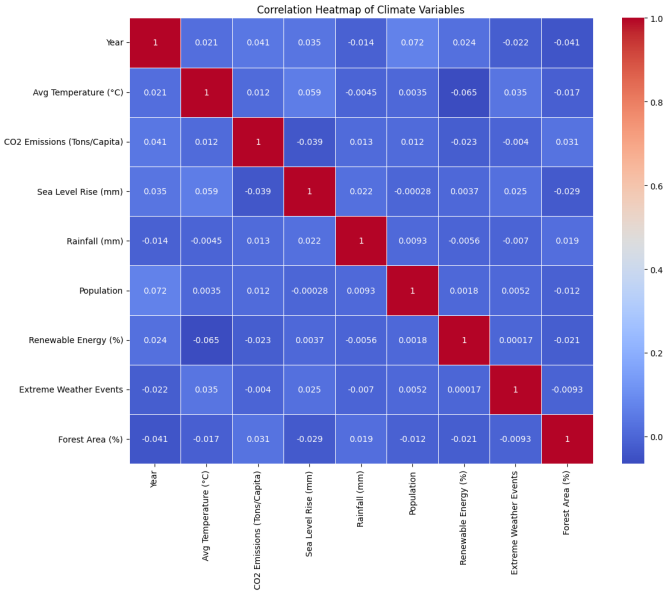


Fig. 2. Correlation heatmap of variables

2) Derived change metrics: We calculated year-over-year changes in key variables (temperature change, sea level change, rainfall change).
3) Development status encoding: Countries were classified as developed or developing.

Data was split with 80% (809 samples) used for training and 20% (176 samples) for testing.

### C. Modeling Approach

We implemented and compared the following models:

- Simple Moving Average (SMA): A baseline approach using the average of the last three years' emissions for a given country
- Linear Models:
  1) Linear Regression
  2) Ridge Regression
  3) Lasso Regression
- Advanced Models:
  1) Random Forest
  2) XGBoost
  3) Support Vector Regression (SVR)

For each advanced model, hyperparameter tuning was performed using either GridSearchCV or RandomizedSearchCV with appropriate cross-validation.
Model performance was evaluated using Mean Squared Error (MSE), $R^2$ score, and Mean Absolute Error (MAE).

## III. Results and Analysis

### A. Model Performance Comparison

All implemented models struggled to achieve positive $R^2$ values, indicating the inherent difficulty in predicting $CO_2$ emissions from the available features. The key performance metrics are summarized in the table below.

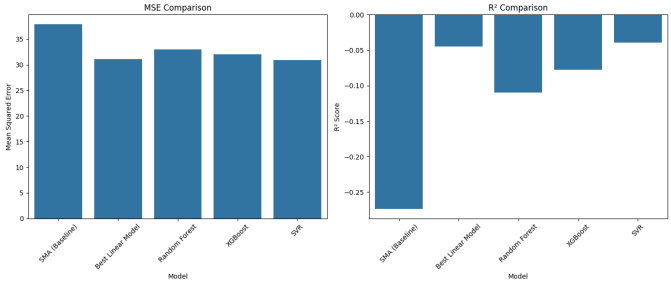


Fig. 3. MSE and $R^2$ values comparison

| Model | MSE | $R^2$ | MAE |
|---|---|---|---|
| SMA (Baseline) | 37.87 | -0.27 | 5.16 |
| Best Linear Model | 31.07 | -0.04 | 4.77 |
| Random Forest | 33.00 | -0.11 | 4.86 |
| XGBoost | 32.04 | -0.08 | 4.79 |
| SVR | 30.92 | -0.04 | 4.75 |

Support Vector Regression (SVR) emerged as the best performer with the lowest MSE (30.92) and MAE (4.75), though its $R^2$ value remained negative (-0.04). The baseline SMA model performed significantly worse than all other approaches. The scatter plots of predicted versus actual values revealed that most models struggled to capture the full range of $CO_2$ emission values, often predicting values clustered around the mean. This pattern is consistent with the negative $R^2$ scores, suggesting limited predictive power beyond simply using the average.

### B. Feature Importance Analysis

Different models identified various features as important predictors:
Linear Model:

- Year
- Average Temperature (°C)
- Sea Level Rise (mm)
- Rainfall (mm)
- Population

Random Forest:

- Renewable Energy (%)
- Population
- Rainfall (mm)
- Previous Forest Area
- Rainfall Change

XGBoost:

- Rainfall Change
- Sea Level Change
- Renewable Energy (%)
- Previous Renewable Energy
- Temperature Change

SVR:

- Development Status (Developing)
- Average Temperature (°C)
- Previous Forest Area
- Forest Area (%)

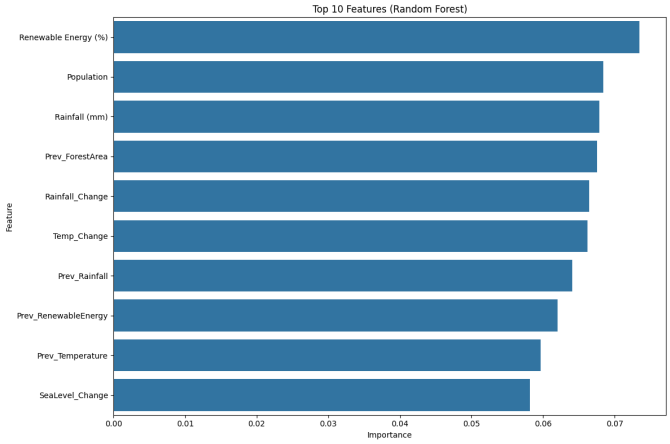


Fig. 4. Random Forest Features

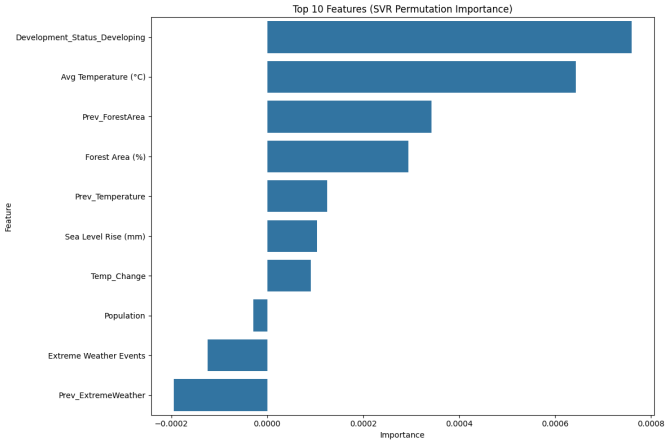


Fig. 6. SVR Features

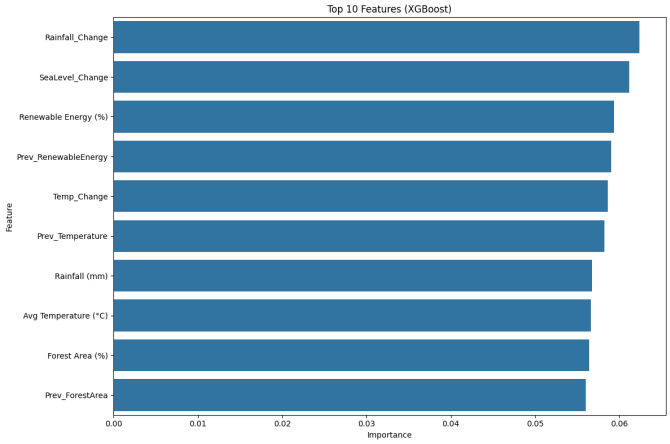


Fig. 5. XGBoost Features

- Previous Temperature

Common important features across multiple models included:

- Renewable Energy (%) - identified by 2 models
- Rainfall (mm) - identified by 2 models
- Previous Forest Area - identified by 2 models
- Average Temperature (°C) - identified by 2 models
- Rainfall Change - identified by 2 models
- Population - identified by 2 models

These overlapping features suggest reliable indicators that consistently influence $CO_2$ emissions across different modeling approaches.

### C. Developed vs Developing Countries Analysis

The best performing model (SVR) showed different prediction accuracy for developed versus developing nations:
Developed Countries:
MSE: 32.56
$R^2$: -0.03
MAE: 4.84
Developing Countries:
MSE: 29.51
$R^2$: -0.05
MAE: 4.67
Interestingly, the model achieved slightly better MSE and MAE metrics for developing countries, though $R^2$ values remained negative for both groups. The SVR model identified development status as one of the most important features, highlighting structural differences in emission patterns between these country groups.

## IV. Discussion

### A. Modeling Challenges

The negative $R^2$ values across all models indicate fundamental challenges in predicting $CO_2$ emissions using the available features. Several factors may contribute to this difficulty:

- Complex socioeconomic factors: Key determinants of emissions like industrial policy, energy infrastructure, and economic structures are not captured in the dataset.
- Policy interventions: Regulatory changes, carbon pricing, and emissions trading systems can cause significant shifts in emission patterns independent of environmental variables.
- Technology adoption: The diffusion of clean technologies follows non-linear patterns that may not correlate strongly with the available indicators.
- Global economic events: Financial crises, pandemic impacts, and trade patterns significantly affect emissions but are not represented in the dataset.
- Timeframe limitations: The relatively short timespan (2000-2023) may not capture long-term climate-emission relationships.

The models' tendency to predict values clustered around the mean suggests they defaulted to safe predictions rather than capturing the full variability in emissions. This highlights the need for additional features or alternative modeling approaches.

### B. Feature Importance Insights

Despite the models' limited predictive power, the feature importance analysis reveals meaningful patterns:

- Renewable Energy Adoption: This factor appeared as an important feature in multiple models, suggesting direct impact on emissions reduction.
- Forest Coverage: Both current and previous year's forest area were identified as important, highlighting the role of natural carbon sinks.
- Climate Dynamics: Rainfall patterns, temperature changes, and sea level metrics were consistently important, indicating complex relationships between climate conditions and emissions.
- Socioeconomic Factors: Population appeared as an important feature in linear and random forest models, reflecting the connection between demographic patterns and emissions.
- Development Status: The SVR model's emphasis on development status suggests structural differences in emission patterns between developed and developing economies.

These findings align with climate science literature that emphasizes the multifaceted nature of emission drivers and the importance of both technological solutions (renewable energy) and natural approaches (forest conservation).

### C. Policy Implications

Based on the feature importance analysis, some possible policy directions emerge:

- Accelerated Renewable Energy Adoption: The consistent importance of renewable energy percentage suggests this remains a high-impact intervention area.
- Forest Conservation and Reforestation: The significance of forest area metrics supports policies focused on protecting and expanding forest coverage.
- Differentiated Approaches by Development Status: The performance differences between developed and developing countries indicate that one-size-fits-all policies may be suboptimal.
- Continuous Monitoring of Key Indicators: Regular tracking of the identified important features could provide early signals of emission trends.
- Comprehensive Data Collection: The models' limited predictive power suggests the need for expanded datasets incorporating economic, technological, and policy variables.

## V. Conclusion

This study demonstrates both the potential and limitations of machine learning approaches in predicting $CO_2$ emissions. While all models struggled to achieve positive $R^2$ values, the analysis identified consistent patterns in feature importance that align with climate science understanding.

The Support Vector Regression model emerged as the best performer, though its limited predictive power suggests that accurate $CO_2$ emission forecasting remains challenging with environmental variables alone. The results highlight the need for expanded datasets that incorporate economic policy, technological adoption, and global event impacts.

The differences observed between developed and developing nations underscore the importance of contextualized climate policies that account for varying socioeconomic conditions.

The identified important features—particularly renewable energy adoption and forest coverage—offer actionable focus areas for emission reduction strategies. Future research should explore incorporating additional socioeconomic variables, longer time horizons, and alternative modeling approaches better suited to the complex, non-linear relationships underlying emission patterns.

## References


[1] IPCC, "Climate Change 2021: The Physical Science Basis," 2021.

[2] Breiman, L., "Random Forests," Machine Learning, vol. 45, pp. 5-32, 2001.

[3] Chen, T., Guestrin, C., "XGBoost: A Scalable Tree Boosting System," 2016.

[4] Smola, A.J., Schölkopf, B., "A tutorial on support vector regression," Statistics and Computing, vol. 14, pp. 199-222, 2004.

[5] Hyndman, R.J., Athanasopoulos, G., "Forecasting: principles and practice," OTexts, 2018.


## VI. Contributions

*Joint Work:*

1) Initial problem formulation and approach planning
2) Final analysis discussions and policy implications
3) Report writing and visualization design
4) Conclusion and future work recommendations

*Anand Shankar Hariharan:*

1) Exploratory Data Analysis (data loading, cleaning, visualizations)
2) Implementation of baseline model (SMA) and Linear regression model
3) Support Vector Regression (SVR) implementation and tuning
4) Feature importance analysis for linear models and SVR
5) Developed vs. Developing countries comparative analysis

*S Mohammed Ashraf:*

1) Data preprocessing and feature engineering (time-lagged features, encoding)
2) Implementation of advanced models (Random Forest, Gradient Boosting)
3) Model evaluation metrics and comparison framework
4) Feature importance analysis for tree-based models
5) Interpretation of results and model performance