Reproducing and Extending Machine Learning Models for Predicting Anthropogenic CO2 Emissions

Aditya Singh

Undergraduate Student (NIT Kurukshetra ’28)

AI & ML Research Enthusiast

singh2709aditya@gmail.com

linkedin.com/in/aditya-singh-63b427321/

September 2025

Abstract

Accurate monitoring of anthropogenic CO_2 emissions is critical for climate mitigation strategies and sustainable development policies. This report presents the reproduction and extension of a previously published research study that modeled CO_2 emissions using machine learning and clustering techniques. The original study focused on China's emission patterns and applied boosting algorithms for predictive modeling. This study not only reproduced the methodology using globally available datasets but also extended it by experimenting with multiple clustering algorithms (such as GMM and Hierarchical Clustering), dimensionality reduction techniques, anomaly detection, temporal analysis, and deriving actionable policy insights. The results provide scalable, interpretable, and data-driven approaches to understanding emission patterns worldwide, contributing to the broader field of environmental AI research.

---

1. Introduction

Anthropogenic CO_2 emissions, primarily driven by industrialization, urbanization, and energy consumption, are among the leading contributors to climate change. As global attention turns toward sustainable development, accurate prediction and analysis of emission trends have become indispensable.

The research paper “A Novel Approach for Predicting Anthropogenic CO_2 Emissions Using Machine Learning Based on Clustering of the CO_2 Concentration” introduced a framework that clustered emission patterns and applied gradient boosting algorithms to forecast emissions for China. Motivated by this study’s methodology, I undertook a reproduction study, validating its methods using open datasets, and an extension study, where I introduced additional algorithms and analyses to generalize findings to a global context.

The goal of this project is to demonstrate how data science can be leveraged to build transparent, scalable, and actionable models that aid climate research and policy design — while showcasing methodological rigor, reproducibility, and innovation.

The complete code, dataset processing steps, and supplementary materials are available on GitHub:

github.com/aditya27singh/ML_Reproduction_and_Extension

---

2. Dataset Overview

Primary Sources

- Our World in Data (OWID) datasets covering global CO_2 emissions, GDP, renewable energy usage, and population.

Features and Indicators

- CO_2 growth, CO_2 growth per capita, GDP growth, renewable growth, renewable share, fossil dependency, cumulative emissions, intensity measures, population growth, log GDP per capita.

These features were engineered to capture the dynamics of emissions and socio-economic factors, enabling meaningful clustering and predictive insights.

---

3. Methodology

3.1 Reproduction Phase

- While the original study focused on boosting-based models like LightGBM, XGBoost etc., the modeling approach of clustering was used in this reproduction.
- Clustering using KMeans with three clusters, replicating the original study's design.
- Dataset adaptations using global data rather than region-specific data, enhancing applicability.
- Evaluation via silhouette scores and centroid analysis to validate clustering patterns.

3.2 Extension Phase

- Inclusion of advanced analytical methods to deepen understanding:
    - Clustering Algorithms: KMeans, Hierarchical Clustering, Gaussian Mixture Models (GMM).
    - Dimensionality Reduction: Principal Component Analysis (PCA) vs non-PCA analysis to assess impact on clustering performance.
    - Anomaly Detection: Identification of outlier countries with abnormal emission trends.

- Temporal Analysis: Evolution of emission clusters across years to detect patterns and shifts.
- Policy Insights: Interpreting cluster characteristics to propose targeted interventions.

3.3 Evaluation Metrics

- Silhouette Score to measure cluster compactness.
- Visual inspection of centroids and cluster distribution.
- Temporal trend graphs to track emission dynamics.
- Analytical reasoning to derive insights for climate policies.

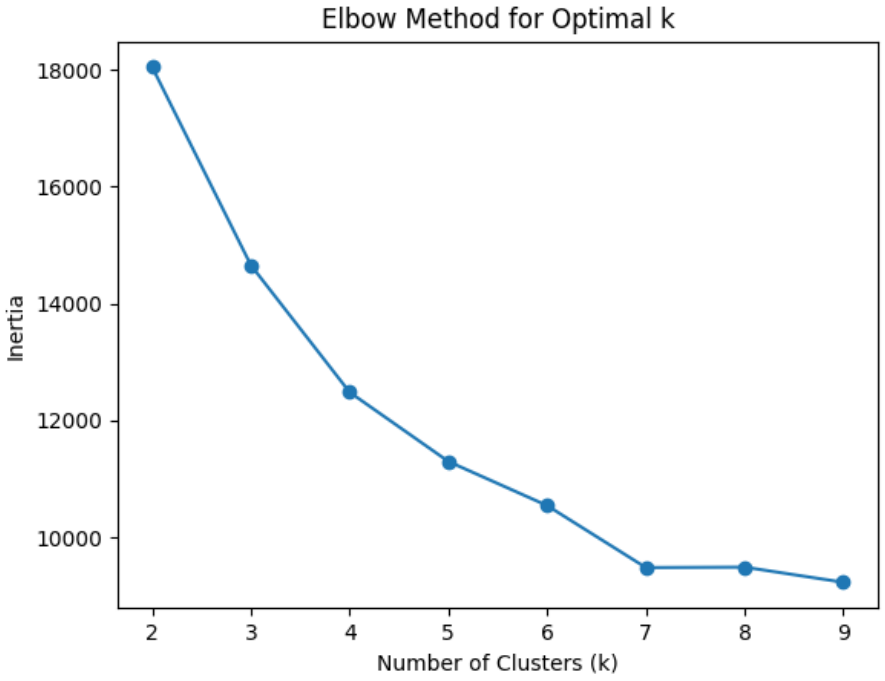
---

4. Results

4.1 Reproduction

4.1.1 Cluster Analysis

The optimal number of clusters was found by using silhouette score and elbow method, and even though the score indicated that k = 2 (0.58) resulted in the most well-separated clusters, the elbow method suggested that k=3 captures additional structure.



The KMeans clustering was performed with taking three clusters, which revealed three distinct country groups:

1. High-emission economies – Large cumulative emissions, moderate renewable adoption.
2. Developed transitional economies – High renewable share, stable emission growth.
3. Mid-industrializing economies – Rapid growth in emissions and energy use.

The computed cluster centroids provide a summary profile of each group, and were instrumental in labelling the clusters as high-emission, transitional, and mid-industrializing economies.

Reproduction:

Features	Cluster 0	Cluster 1	Cluster 2
CO2 growth	0.013	-0.012	0.031
CO2 per capita growth	0.003	-0.018	0.024
GDP growth	0.021	0.019	0.059
Renewable growth	0.116	0.045	0.046
Renewable share	0.046	0.438	0.001
CO2 intensity GDP	0.019	0.001	0.586
Population growth	0.009	0.006	0.007
Fossil dependency	99.954	99.562	99.999
Cumulative CO2	5335.836	849.635	106091.819
Log GDP per capita	9.990	10.502	9.626

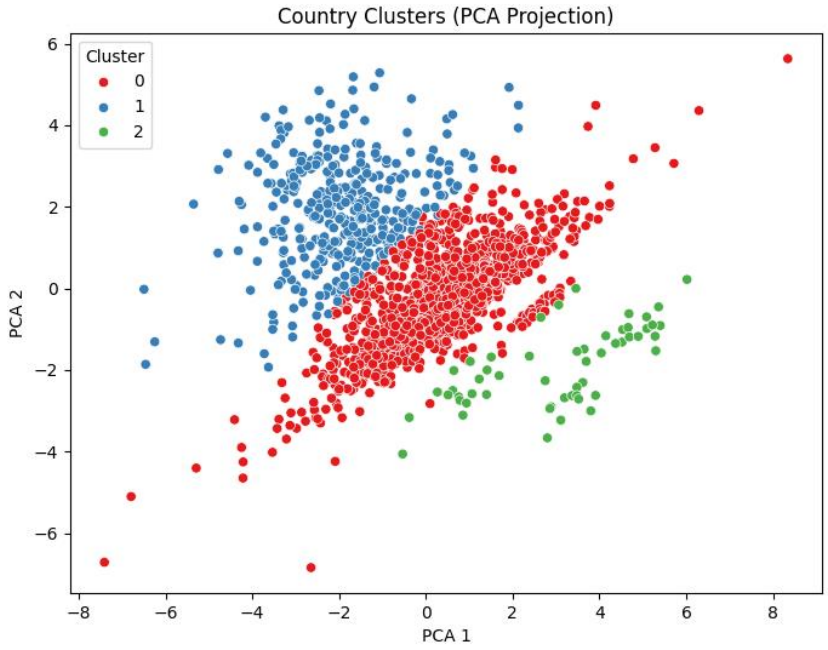
Extension:

Features	Cluster 0	Cluster 1	Cluster 2
CO2 growth	-0.031	-0.008	0.066
CO2 per capita growth	0.038	-0.013	0.053
GDP growth	0.003	0.020	0.044
Renewable growth	0.174	0.036	0.039
Renewable share	0.045	0.457	0.057
CO2 intensity GDP	0.014	0.001	0.065
Population growth	0.007	0.006	0.012
Fossil dependency	99.955	99.543	99.943
Cumulative CO2	8480.591	824.188	8300.132

Log GDP per capita	10.271	10.562	9.619

The centroids presented in the reproduction and extension notebooks differ slightly due to the method used to calculate them. In the reproduction notebook, the centroids are derived directly from the KMeans algorithm's output, while in the extension notebook the centroids are computed by averaging the actual data points within each cluster based on the assigned labels.

For the policy insights, we use the centroid values calculated in the extension notebook.



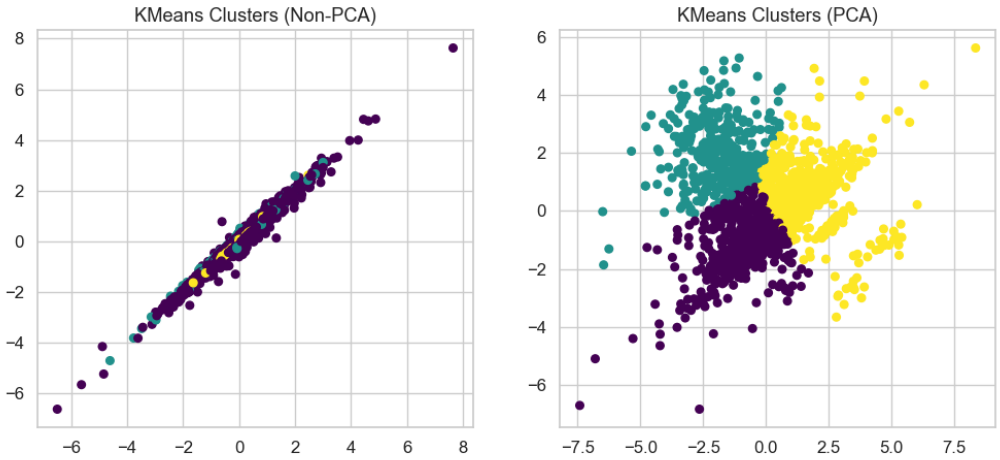
4.2 Extension

4.2.1 PCA vs Non-PCA Comparison for KMeans Clustering

- KMeans clustering was applied to both original data and PCA-reduced data, to gain insights on which method would lead to better-defined, more cohesive clusters, which was estimated by using silhouette score.

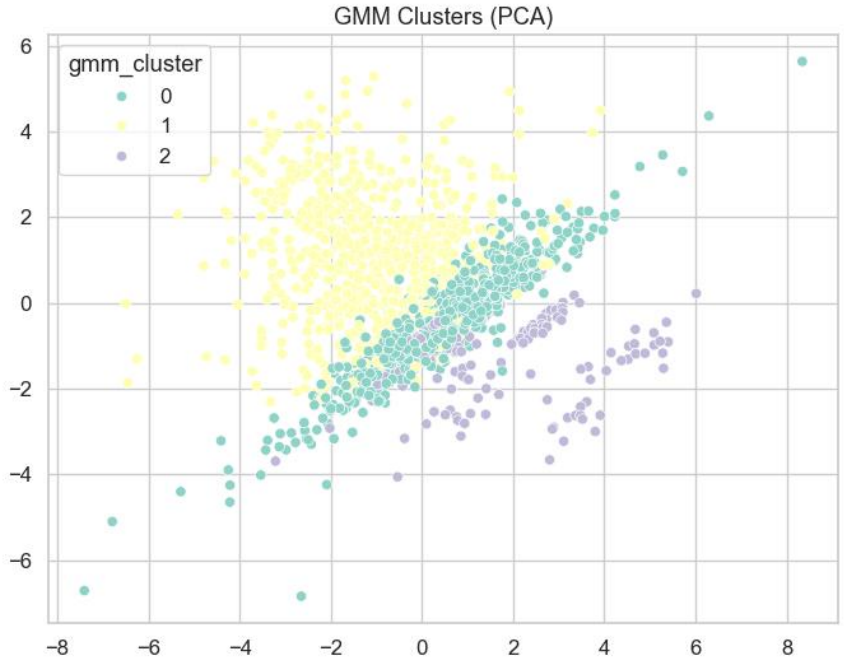
- Results show that silhouette score improvements with PCA reached +0.073, confirming that dimensionality reduction enhanced cluster interpretability without compromising data variance.

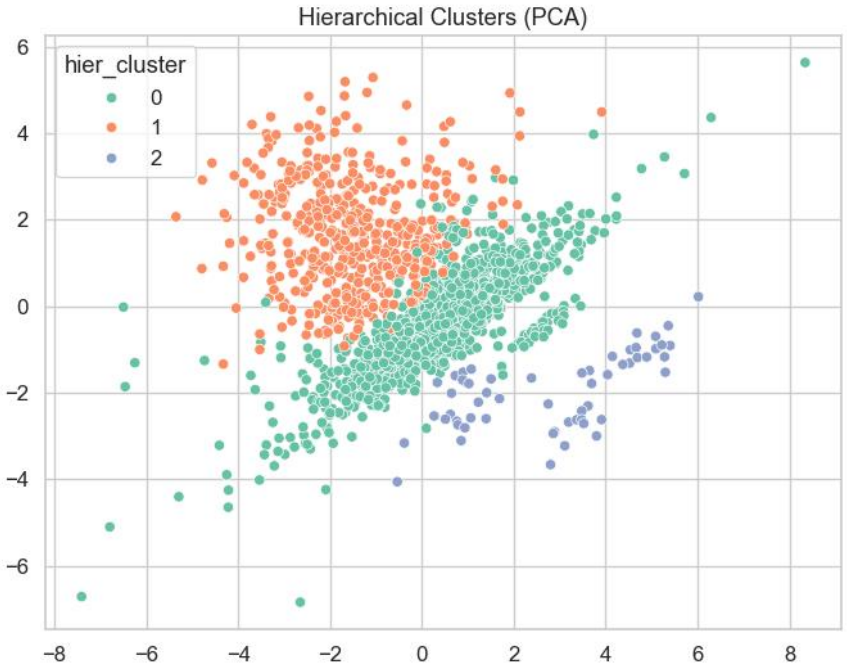
- PCA dimensionality reduction was only applied to KMeans as repeating PCA for Hierarchical Clustering and Gaussian Mixture Models would not provide the same clarity in visualization or interpretability.



4.2.2 Comparative Algorithm Insights

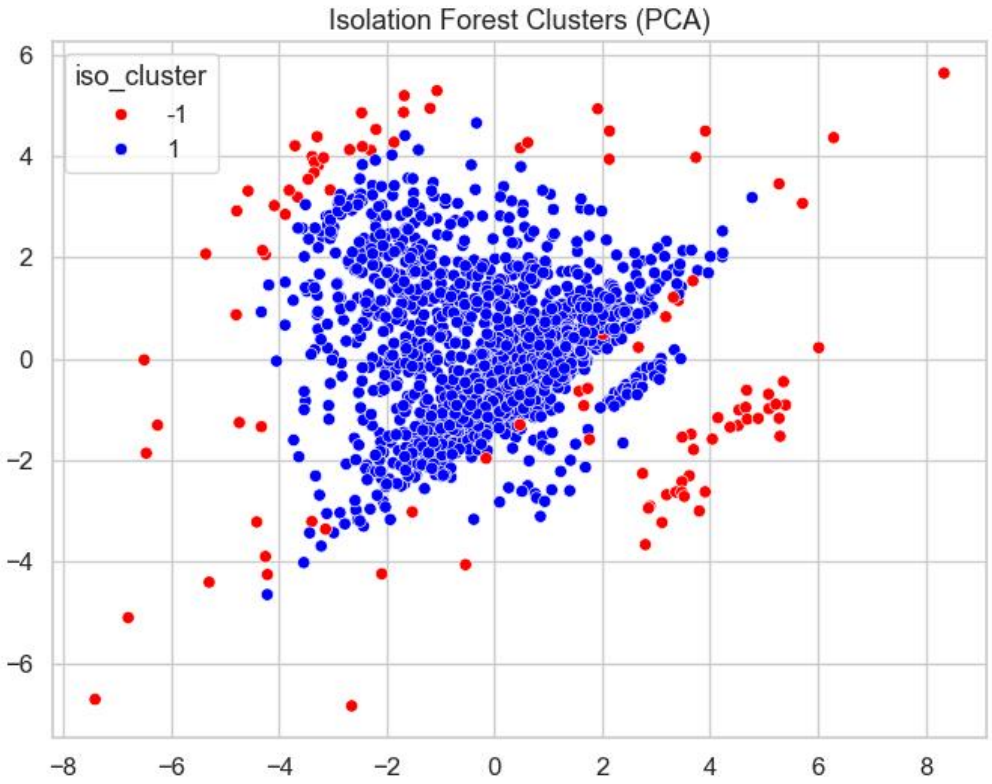
- Silhouette score was calculated for all three clustering models: KMeans (0.38), Hierarchical Clustering (0.25), and Gaussian Mixture Model (0.08).
- Hierarchical Clustering uncovered relationships between countries with similar socio-economic structures.
- GMM provided probabilistic cluster memberships, highlighting overlaps and uncertainties in emission behaviors
- The silhouette scores show us that KMeans achieved the most stable and interpretable clusters, as it has the highest silhouette score out of all three models.





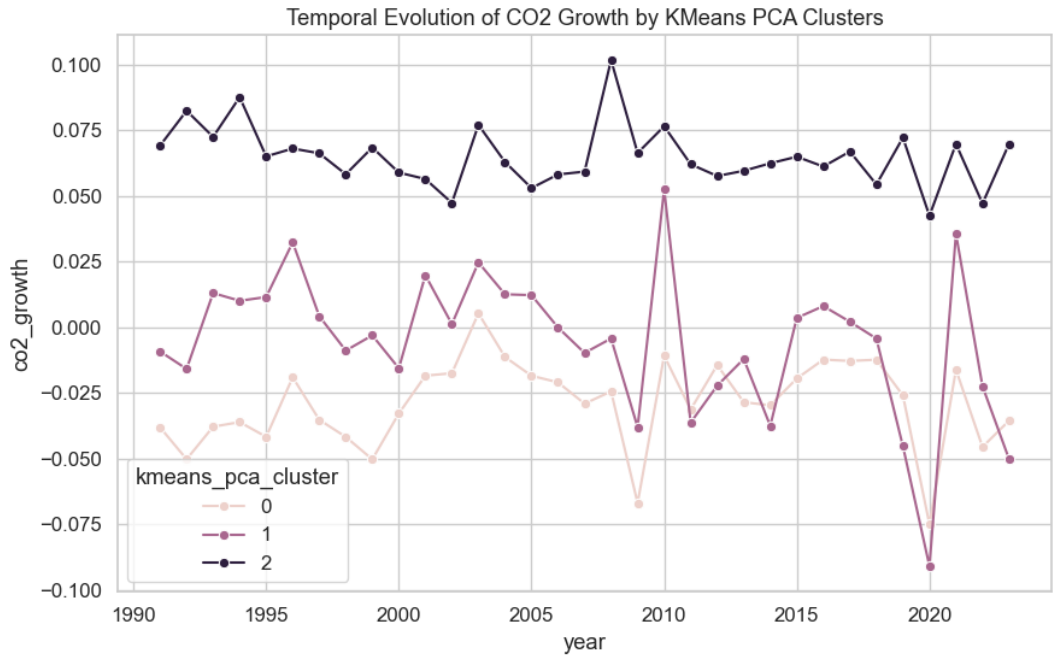
4.2.3 Anomaly Detection Findings

Outlier nations with sudden emission changes were identified, indicating regions needing urgent environmental assessments or policy intervention. These sudden changes highlighted outlier years (e.g., COVID-19 drop in 2020)



4.2.4 Temporal Trends

Emission patterns evolved distinctly across clusters, with some mid-industrializing economies showing increasing dependence on fossil fuels, while transitional economies demonstrated renewable adoption strategies.



4.5 Policy Implications

- High-emission countries need targeted interventions focusing on energy diversification and urban planning. Phasing out fossil fuels, introducing carbon pricing mechanisms and other stringent policies are needed to prevent locking into high-carbon pathways.
- Transitional economies should prioritize accelerating renewable energy investment and energy efficiency policies, as they have the infrastructure and emissions profile of industrialized nations but are lagging in sustainable transition.
- Mid-industrializing nations require support frameworks for renewable energy deployment and climate risk mitigation. These economies demonstrate strong progress. Policies should focus on maintaining emissions reductions and serving as models for sustainable development pathways

---

5. Conclusion

This research successfully reproduced a published study while extending its scope through global datasets and methodological innovations. The project

bridges machine learning and climate science, offering transparent and interpretable insights into emission patterns.

Key achievements:

- Validated existing methodologies with open data.
- Generalized findings for broader applications.
- Extended the paper to various clustering models, and other machine learning aspects.
- Demonstrated reproducibility, rigor, and research-driven thinking — crucial for academic and industry collaborations.

---

6. Future Work

Although this study does not extend beyond the current scope, several promising avenues for future research exist:

1. Integrating additional datasets
2. Refining temporal modeling and
3. Enhancing interpretability through advanced methods

These could improve the accuracy and applicability of CO2 emission predictions.

---

7. Acknowledgments

I am grateful to the authors Zhanghui Li, Hao Song, Lipling Lei, Mengya Sheng, Kaiyun Guo, and Shaoqin Zhang of the original research paper *"A Novel Approach for Predicting Anthropogenic CO2 Emissions Using Machine Learning Based on Clustering of the CO2 Concentration",* originally published in *Atmosphere, 2024* for providing inspiration and methodological clarity, and to the global data providers at Our World in Data for enabling reproducibility and transparency in environmental research.