Analyzing CO_2 Emissions to Understand Climate Trends and Predict Future Impact

1. Introduction

This report outlines a data science project focused on analyzing global carbon dioxide (CO2) emissions, their relationship with economic indicators, and their impact on global temperature rise. The project utilizes machine learning and time-series forecasting techniques to analyze emission patterns, model temperature changes, and forecast future emission trajectories.

2. Team

This is an individual project completed by:
Aqdas Bibi

3. Problem

The increasing concentration of greenhouse gases, especially CO_2, is a key contributor to global warming. Understanding emission trends and their links with temperature rise is essential for climate policy. This project examines the relationship between economic growth, energy consumption, and emissions in the context of climate change, and aims to predict future emissions based on historical patterns.

4. Objective

This project aims to explore the rise of CO_2 emissions globally and assess their relationship with temperature changes over time. The project involves:

- Cleaning and transforming a real-world CO_2 dataset
- Visualizing emission trends by country
- Analyzing feature correlations
- Predicting temperature change using regression models

5. How is it related?

The techniques used in this project, such as regression, clustering, and time series forecasting, are all part of modern data science curricula. The project combines environmental science with statistical modeling, providing practical experience in feature engineering, visualization, and predictive analytics.

6. Dataset

The dataset used in this project is sourced from **Our World in Data – CO_2 and Greenhouse Gas Emissions**. It provides comprehensive, country-level data spanning from **1750 to 2022**, although this analysis focuses on records from **1960 onward** to capture the modern industrial period. The dataset includes key variables such as country, year, CO_2, CO_2_per_capita, GDP, population, energy_per_capita, and temperature_change_from_CO_2. It also breaks down emissions by source, including cement_co_2, coal_co_2, gas_co_2, and oil_co_2. This rich set of features enables a detailed investigation of how economic development and energy use correlate with rising carbon emissions and global temperature change.

7. Proposed Methodology

The project followed a structured pipeline that included data preparation, analysis, modeling, and interpretation. The methodology is outlined as follows:

1. Selected relevant variables such as CO_2 emissions, GDP, population, and energy use.
2. Filtered the dataset to include records from **1960 onward**, focusing on the modern industrial era.
3. Removed non-country aggregate records (e.g., World, Asia, Europe) to ensure country-level consistency.
4. Retained rows with at least **8 non-null values** to maintain data integrity.
5. Dropped rows with missing values in **key target and predictor variables**, particularly CO_2 and temperature_change_from_CO_2.
6. Perform exploratory data analysis and feature engineering.
7. Apply **regression** models (Ridge, Random Forest) to predict temperature rise
   - Performed **data cleaning** and handled missing values in model features.
   - Split the data into **training (80%)** and **testing (20%)** sets.
   - Trained the regression model and evaluated it using:
     - **R^2 Score**: Measuring the proportion of variance explained.
     - **RMSE (Root Mean Squared Error)**: Measuring the average prediction error.
8. Use K-Means clustering to group countries by emissions and economic indicators.
9. Forecast CO_2 emissions using Facebook Prophet time-series model.
10. Visualize results using Seaborn and Plotly.

Reset the index for the cleaned DataFrame.

8. Expected Outcomes

- Clear evidence of increasing CO_2 levels in industrialized and developing countries
- Temperature change shows a measurable response to CO_2 emissions
- Regression modeling shows predictive power using standard economic and environmental variables
- The correlation matrix supports the expected theoretical relationships
- Reliable CO_2 forecasts for the next 10 years

9. Modules

- Data Preprocessing and Cleaning
- Feature Engineering
- Regression Modeling
- Clustering
- Time Series Forecasting
- Data Visualization

10. Limitations & Future Work

- The regression model used only a few engineered features — more variables could improve accuracy.
- Temperature data was modeled (not measured), which may affect real-world interpretation.
- Forecasting CO_2 or temperature using **Prophet** or **ARIMA** models would enhance the project.
- Clustering countries based on emission profiles could add another dimension to the analysis.