**Problem Statement**

This project focuses on analysing country-level data and building machine learning models to forecast CO₂ emissions based on a range of national indicators. It utilizes the publicly available *Climate Change Data* provided by the World Bank Group, which includes comprehensive data spanning multiple countries and years. The dataset encompasses a wide array of features such as:

* **Countries covered:** A majority of nations across the globe
* **Time frame:** Data from 1990 to 2011
* **Greenhouse gas emissions:** Including CO₂, CH₄, N₂O, among others
* **Demographic details:** Total population, urban population, population growth rate, etc.
* **Economic indicators:** GDP, GNI, foreign investments, etc.
* **Agricultural and land-related data:** Cereal yield, agricultural area, protected land zones, etc.
* **Climate metrics:** Precipitation levels, natural disaster counts, etc.
* **Energy consumption**
* **Healthcare-related statistics:** Number of medical professionals, etc.

The project workflow is divided into two major phases:

**Phase 1: Data Preprocessing and Structuring**

This phase is documented in a dedicated Jupyter Notebook (.ipynb) along with a corresponding PDF report. It includes:

* **Introduction:** Summary of the notebook and the project context, including data source information
* **Initial Setup:** Importing necessary libraries and loading the dataset
* **Dataset Overview:** A global perspective on the data available
* **Project Objectives:** Outlining the goals of the analysis
* **Data Cleaning Tasks:**
* Addressing missing data
* Converting non-numeric values into numeric formats
* Renaming columns for clarity
* Eliminating rows and columns lacking data
* **Data Transformation Procedures:**
* Reshaping variables using data melting techniques
* Merging different data subsets into a unified format
* **Handling Missing Entries:**
* Identifying and filtering out missing values with minimal data loss
* **Final Step:** Exporting the cleaned and structured dataset for further analysis

**Phase 2: Data Exploration and Visualization**

This phase is documented in a dedicated Jupyter Notebook (.ipynb) along with a corresponding PDF report. It includes:

* **Introduction:** Notebook setup with data source information and imported libraries
* **Global Data Overview:** A high-level perspective on the dataset structure and contents
* **Feature/Column Definitions:** Description of abbreviations and measurement units used in the dataset
* **Hypothesis Statement:** Defining the core hypothesis to be tested during analysis
* **Feature Engineering:**
* Overview of relevant features selected for analysis
* Derivation of additional key features for improved insights
* Removal of features deemed irrelevant or redundant
* **Data Visualization Preparation:**
* Creating visualizations to understand feature relationships
* Generating global views and focused plots of important dependencies
* Visualization types include correlation matrix heatmaps, scatterplots, and histograms
* **Outlier Detection:** Identifying and examining abnormal data points that may impact analysis
* **Discussion:** Interpreting observed dependencies and trends from the visualizations
* **Conclusions:** Summarizing key findings and how they relate to the initial hypothesis

**Phase 3: Model Development and Evaluation**

This phase is documented in a dedicated Jupyter Notebook (.ipynb) along with a corresponding PDF report. It includes:

* **Introduction:** Summary of the notebook and project context, with notes on the data source
* **Initial Setup:** Importing necessary libraries, loading the dataset, and addressing randomness in algorithm behavior
* **Dataset Overview:** A concise look at the dataset and its structure
* **Feature Definitions:** Overview of abbreviations used for dataset features and columns
* **Hypothesis Statement:** Definition of the core hypothesis to be tested using the model
* **Variable Selection:**
* Identification of dependent (target) and independent (predictor) variables
* Splitting the dataset into training and testing subsets for evaluation
* **Feature Selection Procedures:**
* Applying recursive feature elimination (RFE)
* Integrating cross-validation to ensure robust feature relevance
* **Model Tuning:**
* Hyperparameter optimization of a Random Forest model using cross-validation
* **Model Training and Evaluation:**
* Training the model with optimal hyperparameters using the training subset
* Performance assessment through cross-validation
* **Test Set Validation:**
* Validating model performance on the previously unseen test data subset
* **Conclusions:** Final remarks summarizing model performance and its alignment with the hypothesis