1 Background and context

Real world machine learning applications involve many aspects, often, model building is a relatively small piece of the overall application. Although the project mainly focuses on machine learning and data analytics aspects, the overall aim is to provide hands-on practice with machine learning tehniques, while at the same time encouraging thinking through the problem in a holistic manner. An important skillset is to ask the right questions which drives the analysis in practically useful directions.

Climate change is a current topic of enormous significance. It is well known that emissions of greenhouse gases such as CO2 are a significant cause of global warming. The dataset to be used for this project consists of key climate change indicators collected across different countries from the year 2000 to 2024. The indicators include average temperature, CO2 emissions, sea-level rise, rainfall patterns and so on. The data could be used to analyze trends, correlations, and anomalies.

2 Project description

2.1 Part 1) Model building and evaluation

(A) **Exploratory Data Analysis** Understanding the data and the domain is a very important aspect in any real world machine learning application. Typically, this is done via simple summaries and visualizations. Commonly used visualizations include scatter and box plots. A line plot is a simple visualization for time series data (i.e. where time is the independent variable and there are one or more dependent variables), A bubble chart is another useful visualization to represent multiple variables in a time series, usually the x-axis represents time, the y-axis represents one variable and the size of the bubble represents a third variable. Include suitable visualizations which would be useful for illustrations or deriving insights.

(B) **Build a predictive model to predict per capita CO2 emissions** Compare predictions of CO2 per capita emissions using the following approaches.

- Baseline : Compute CO2 emissions using a **Simple Moving Average**
- A **Linear regression** model using the other attributes as predictors. While you may choose to exclude one or more attributes, it should be backed up with a clear justification.
- A more sophisticated model which should ideally be much better than the above baseline models. Candidates include ensemble models such as random forests, Gradient Boosting trees etc.

2.2 Part 2) Analysis

(A) **Which are the most important predictors?**
This can be approached in multiple ways via machine learning, although there is no perfect answer. For instance, the coefficients in Multiple Linear Regression can be used to assess the impact of predictors (note that for non-standardized variables, the coefficients depend on units of the variables). Also, note that the most important predictors may not necessarily be the most actionable ones or, they may be the effects rather than the cause.

(B) **Analyze the effects of various indicators for developing and developed countries separately** Suggest how the analyses above can be used towards framing a policy to cut down global emissions (both per capita as well as overall CO2 emissions). It is possible that the data provided may not be sufficient, if so suggest what additional information, in your opinion, would be helpful in this regard.