# Project:
# CO2 Emissions Using LSTM

## Faculty of Computer Science and Info. Systems
## October 6 University

## Team Members:

| Name | ID |
|---|---|
| Ahmed Ismail El Sayed El Sayed Azzam | 202201998 |
| Ahmed Samy Abdel Latif | 202202718 |

## Under the Supervision of:
## Dr. Doaa Bliedy

# 1. Introduction

The escalation of global warming and climate change stands as one of the defining challenges of the 21st century, driven primarily by the accumulation of greenhouse gases in the atmosphere. Among these, Carbon Dioxide (CO2) is the most significant contributor, originating from diverse economic activities ranging from industrial production to domestic energy consumption. As nations strive to meet sustainability targets and mitigate environmental impact, the ability to monitor historical emission patterns and accurately forecast future trends has become indispensable for researchers and policymakers.

This project leverages a comprehensive CO2 Emissions Dataset, which provides granular daily records of emissions across multiple countries and sectors, including Power, Industry, Ground Transport, Residential, and Aviation. The primary objective of this study is to transform raw emission data into actionable insights through rigorous Exploratory Data Analysis (EDA) and advanced Time-Series Forecasting.

The project workflow begins with data preprocessing and visualization, offering a comparative look at how major global economies such as the United States, China, India, and European nations have contributed to emissions over time. By visualizing sector specific trends, we aim to identify seasonal patterns, anomalies, and the distinct carbon footprints of different industries. Following the analysis, the project implements a Deep Learning approach using a Long Short-Term Memory (LSTM) neural network. LSTMs are uniquely suited for this task due to their ability to learn long-term dependencies in sequential data. The resulting model is designed to predict daily CO2 emissions across six distinct sectors simultaneously, providing a robust tool for environmental monitoring and future planning.

## 2. Dataset Description

The analysis utilizes a comprehensive time-series dataset recording daily Carbon Dioxide (CO2) emissions. This dataset provides a granular view of the environmental footprint across 14 distinct countries, broken down into 6 specific economic sectors. The emission values are standardized and measured in $MtCO_2$/day (Million Tonnes of Carbon Dioxide per day).

**Dataset Overview**

The raw data is structured in a "long" format, where every individual record represents the emissions for a specific sector in a specific country on a single day.

- Total Records: 135,408 entries.
- Total Columns: 5 features.

The dataset consists of the following five variables:

- **country** (Categorical): The name of the nation associated with the record (e.g., Brazil, China, US, India).

- **date** (Object/Date): The specific calendar date of the observation (formatted as DD/MM/YYYY).

- **sector** (Categorical): The economic category generating the emissions. The dataset covers six distinct sectors:
  - Power
  - Industry
  - Ground Transport
  - Residential
  - Domestic Aviation
  - International Aviation

- **value** (Numerical): The quantity of CO2 emitted, measured in **$MtCO_2$/day**.

- **timestamp** (Numerical): A Unix timestamp representation of the date, used for numerical sorting or processing.

# 3. Data Preparation and Feature Engineering

To ensure the dataset was suitable for training a Long Short-Term Memory (LSTM) neural network, a rigorous preparation pipeline was implemented. This process transformed the raw, categorical records into a normalized, multivariate time-series format.

## 3. 1 Data Cleaning and Formatting

Datetime Conversion: The date column, originally read as a string object, was converted into datetime objects using (pd.to_datetime) This ensured chronological accuracy and allowed for proper sorting.

Quality Check: A null-value assessment was conducted to ensure data integrity. The dataset was found to be complete with no missing values (NaN) across all 135,408 records.

## 3.2 Data Transformation (Pivoting)

The raw data was structured in a "long" format (one row per sector). To capture the interdependencies between different economic sectors, the data was reshaped into a "wide" format using a pivot operation.

Structure Change**:** The unique values in the sector column were transformed into individual features.

Result: The resulting dataframe consists of a single row per Country/Date combination, with six columns representing the emission values for specific sectors (e.g., Power, Industry, Transport). This step was crucial for enabling the model to learn multi-sector correlations simultaneously.

## 3. 3 Feature Encoding

Since neural networks require numerical input, categorical variables were encoded:

Country Encoding: The country column was transformed using (LabelEncoder) This assigned a unique integer to each of the 14 nations (e.g., Brazil=0, China=1), creating a new feature country_enc. This allows the model to differentiate emission patterns based on geographic regions.

## 3. 4 Temporal Train-Test Split

To prevent "data leakage" where a model inadvertently learns from future data random shuffling was avoided. Instead, a chronological split was applied specifically for each country to maintain temporal continuity:

- Training Set: The first 70% of the timeline.
- Validation Set: The subsequent 15% of the timeline.
- Test Set: The final 15% of the timeline.

## 3.5 Feature Scaling (Normalization)

LSTMs are sensitive to the scale of input data. Large values (like emissions in millions of tonnes) can cause instability during training.

**Min-Max Scaling:** Both the input features (X) and target variables (y) were scaled to a range of **[0, 1]** using MinMaxScaler.

**Scalers:** Two separate scalers were maintained one for features and one for targets to ensure accurate inverse transformation of predictions later.

## 3.6 Time-Series Windowing (Sliding Window)

To forecast future emissions, the data was restructured into sequential windows using a Sliding Window technique.

**Look-back Period (Time Steps):** A window size of 12 days was selected.

**Logic:** The model utilizes data from days $t - 12$ to $t - 1$ to predict emissions for day t.

**Input Shape:** The final processed data for the LSTM input was shaped as (Samples, 12, Features), representing the number of samples, the 12-day look-back period, and the 7 features (6 sectors + 1 country code).

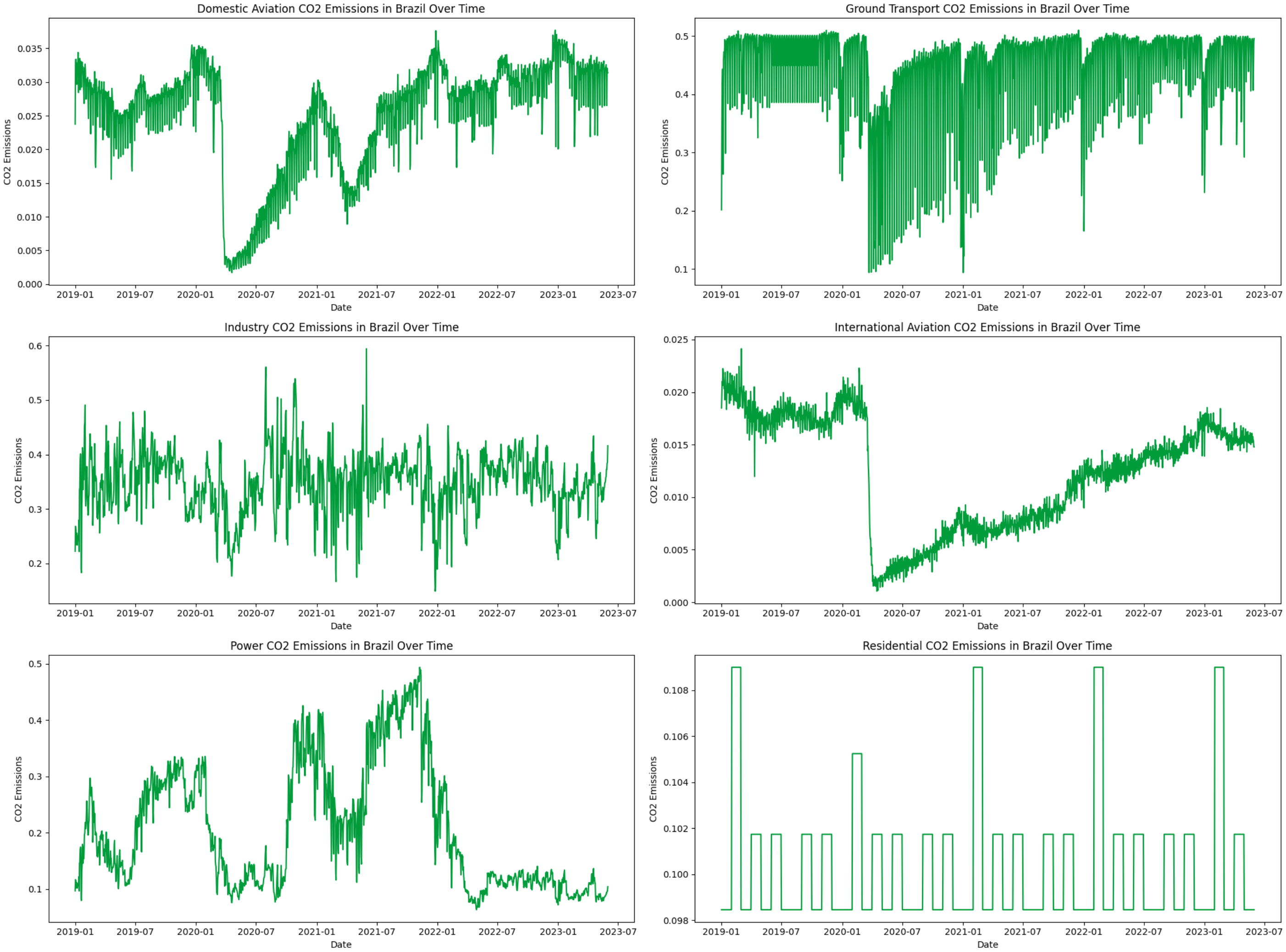
# 4. Exploratory Data Analysis and Visualizations

To understand the temporal dynamics and distinct emission patterns of different economies, we conducted a detailed visual analysis. The primary goal was to observe how Carbon Dioxide (CO2) emissions have evolved from 2019 to 2023 across six key sectors: Domestic Aviation, Ground Transport, Industry, International Aviation, Power, and Residential.

## 4. 1 Country-Specific Sector Analysis

For each of the 14 nations in the dataset, we generated multi-panel time-series plots. To ensure clarity given the vast difference in scale between sectors (e.g., the high volume of "Power" vs. the lower volume of "Domestic Aviation") we utilized a 3x2 subplot grid layout. This approach allows for an independent Y-axis for each sector, ensuring that granular trends are not obscured by high-emission sectors.

## 4.2 Case Study: Brazil



As a representative example, the visualization for Brazil reveals several critical insights into the country's carbon footprint:

- Transport Volatility: The "Ground Transport", "Domestic Aviation" and "Intrnational Aviation" sectors show significant fluctuations. A sharp decline is evident in early 2020, corresponding to the onset of the COVID-19 pandemic and subsequent travel restrictions.
- Power Sector Stability: Unlike aviation, the "Power" sector demonstrates a more consistent baseline, though distinct spikes suggest periods of high energy demand or reliance on carbon-intensive energy sources during droughts (affecting hydroelectric capacity).
- Seasonal Trends: The "Residential" sector displays periodic wave-like patterns, indicative of seasonal energy consumption for cooling or heating.

### 4. 3. Global Observations (All Countries)

By replicating this visualization strategy for all nations (including China, US, India, and European countries), several universal and region-specific trends emerged:

- The "COVID Dip": Across almost every nation, a massive, synchronized drop in emissions occurred in the first quarter of 2020. This was most pronounced in the International Aviation and Ground Transport sectors, visualizing the global economic standstill.
- Recovery Patterns: The plots highlight varying speeds of economic recovery. While some nations returned to pre-pandemic emission levels by 2021, others showed a slower, more gradual return.
- Seasonality: Northern hemisphere countries (like Russia, Germany, and the US) exhibit strong seasonality in the Residential sector, with sharp peaks during winter months due to heating demands, contrasting with the patterns observed in equatorial or southern hemisphere nations.

This visual exploration confirms that CO2 emissions are not linear but are heavily influenced by external global events, seasonal cycles, and specific national energy policies. These insights validate the need for a non-linear predictive model, such as LSTM, to capture these complex dependencies.

# 5. Methodology

To address the complexity of predicting multi-sector CO2 emissions, we employed a Deep Learning approach utilizing Long Short-Term Memory (LSTM) networks. LSTMs are a specialized variant of Recurrent Neural Networks (RNNs) capable of learning long-term dependencies in sequential data, making them ideal for time-series forecasting where past trends strongly influence future outcomes.

## 5.1 Model Architecture

- Input Layer: Receives a sliding window of data with a shape of (12, 7), representing the past 12 days of data across 7 features (6 sectors + encoded country).
- First LSTM Layer: A layer with 128 units and return_sequences=True. This layer captures high-level temporal patterns and passes the sequence to the next layer.
- Dropout Layer (20%): A regularization technique used to randomly ignore 20% of the neurons during training to prevent overfitting.
- Second LSTM Layer: A layer with 64 units. This layer condenses the temporal information into a comprehensive feature vector.
- Second Dropout Layer (20%): Further regularization to ensure model generalization.
- Output Dense Layer: A fully connected layer with 6 units and a ReLU (Rectified Linear Unit) activation function.
    - Significance: This multi-output design allows the model to predict emissions for all 6 sectors simultaneously (Domestic Aviation, Ground Transport, Industry, International Aviation, Power, Residential). The ReLU activation ensures that predicted emission values are non-negative.

## 5.2 Training Configuration

- Optimizer: The Adam optimizer was selected for its adaptive learning rate capabilities.
- Loss Function: Mean Squared Error (MSE) was used as the objective function to minimize the squared differences between predicted and actual emissions.

Hyperparameters:

- Epochs: 30 (iterations over the entire dataset).
- Batch Size: 32 (samples processed before updating internal model parameters).
- Validation: Performance was monitored against a 15% validation set during training to track convergence.

## 5.3 MLOps and Experiment Tracking

To ensure reproducibility and maintain a rigorous record of model performance, we implemented an MLOps pipeline using MLflow. This framework allowed us to manage the machine learning lifecycle systematically:

- Metric Logging: The model’s key performance indicators MAE, MSE, RMSE, and $R^2$ were automatically logged for every experimental run. This enabled a precise quantitative comparison between different iterations of the model.
- Artifact Management: Visualizations, including scatter plots of "Predicted vs. True" values and sector-specific line plots, were saved as artifacts. This provided a qualitative audit trail, allowing us to visually confirm the model's alignment with ground truth data.
- Model Versioning: The trained LSTM model was logged with an inferred signature (defining the specific input and output schema). This ensures that the model can be loaded and deployed in future environments without compatibility issues regarding data shape or types.

## 5.4 Deployment Interface (Streamlit)

To bridge the gap between complex deep learning architectures and practical usability, we developed an interactive graphical user interface (GUI) using Streamlit.

- Functionality: The application serves as a front-end dashboard where users can select a specific Country and a future Date.
- Real-Time Inference: Upon user input, the system loads the trained LSTM model and the feature scalers in the backend to generate real-time CO2 emission forecasts for all six sectors.
- Visualization: The interface presents the output not just as raw numbers, but as intuitive metrics ($MtCO_2$/day), making the insights accessible to non-technical stakeholders such as policymakers or environmental researchers.

# 6. Results and Discussion

The Long Short-Term Memory (LSTM) network demonstrated exceptional efficacy in modeling the complex, non-linear dynamics of global CO2 emissions. The model was evaluated on a dedicated Test Set (the final 15% of the timeline), yielding high-accuracy forecasts across all six economic sectors.

## 6.1 Quantitative Performance

The model's performance was rigorously assessed using standard regression metrics. The high $R^2$ score indicates that the model successfully captured the underlying patterns, seasonality, and trends of the data.

- $R^2$ Score (Coefficient of Determination): 0.9889
- Insight: The model explains approximately 98.89% of the variance in the test data. This near-perfect score confirms that the LSTM architecture effectively learned the temporal dependencies of emissions.
- RMSE (Root Mean Squared Error): 0.6288
- Insight: On average, the model's predictions deviate from the actual emission values by approximately 0.63 $MtCO_2$/day. Given the scale of global emissions (which can range into the thousands), this error margin is negligible.
- MAE (Mean Absolute Error): 0.2318
- Insight: The low absolute error highlights the model's precision in daily forecasting.
- MSE (Mean Squared Error): 0.3954

## 6.2 Visual Analysis (MLflow Evaluation)

Using MLflow for experiment tracking, we generated visual artifacts to qualitatively assess model fit.

Predicted vs. True Scatter Plot: The scatter plot comparing predicted values against ground truth reveals a tight clustering around the diagonal (y=x) line. This indicates a strong correlation and suggests that the model does not suffer from significant bias (i.e., it does not consistently over-predict or under-predict).

Sector-Specific Trajectories:Line plots generated for individual sectors (Domestic Aviation, Ground Transport, Industry, etc.) demonstrate the model's ability to distinguish between different emission behaviors:

- It accurately reproduced the highly seasonal patterns of the Residential sector.
- It captured the volatility and rapid changes in the Ground Transport and Aviation sectors.
- It smoothed out noise while retaining the general trend for stable sectors like Power.

### 6.3 Deployment and Practical Application (Streamlit)

The deployment of the model via a Streamlit GUI successfully transitioned the project from theoretical analysis to a practical tool.

- Real-Time Inference: The application successfully integrated the saved LSTM model and scalers (scaler_X.pkl, scaler_y.pkl).
- Scenario Testing: In a test case for Brazil on June 1, 2023, the system generated distinct, plausible predictions for all sectors (e.g., predicting ~0.579 $MtCO_2$/day for Ground Transport).
- Usability: The interface allows non-technical users to query specific dates and countries, providing immediate actionable data without interacting with the underlying code.

## 7. Conclusion

This project successfully addresses the critical challenge of monitoring and forecasting global Carbon Dioxide (CO2) emissions through the application of advanced Deep Learning techniques. By leveraging a comprehensive dataset spanning 14 countries and 6 economic sectors, we developed a robust analytical framework capable of deciphering complex, non-linear trends in environmental data.

**Key Achievements:**

- **Data Transformation:** We successfully converted raw, transactional emission records into a structured multivariate time-series. This pivoting strategy allowed for the simultaneous analysis of distinct sectors such as Power, Industry, and Transport capturing the interdependencies between different economic activities.
- **Model Precision:** The implementation of the Long Short-Term Memory (LSTM) neural network proved highly effective. With an $R^2$ score of ~0.9889 and a low Mean Squared Error (MSE) of 0.3954, the model demonstrated an exceptional ability to learn from historical sequences and predict future emissions with high fidelity.
- **Holistic MLOps Integration:** The project moved beyond simple model training by incorporating MLflow for experiment tracking and reproducibility. This ensured that every metric, parameter, and visual artifact was documented, providing a rigorous audit trail for the model's performance.
- **Practical Usability:** The development of the Streamlit GUI transformed the predictive model into an accessible tool. By allowing users to generate real-time forecasts for specific dates and countries, the project bridges the gap between complex data science and actionable environmental insight.