

# IE 6318 Project Proposal

## Analysis and Forecasting of UK Greenhouse Gas Emissions: A Data Mining Approach

### Abstract

This project will conduct a comprehensive data mining analysis of the UK's territorial greenhouse gas emissions from 1990 to 2023. The primary goal is to identify key trends, uncover the primary drivers of emissions, and develop a predictive model to forecast future emission levels. By analyzing emissions by sector, gas, and fuel type, this study aims to provide data-driven insights into the effectiveness of past and present environmental policies. The core data mining problems are to forecast future greenhouse gas emissions through time-series analysis and to segment emission sources using clustering techniques. This work is meaningful as it addresses the critical global challenge of climate change, offering a quantitative basis for evaluating the UK's progress towards its climate targets and informing future policy decisions.

### Project Plan

1. **Dataset Description:** The project will utilize the "Final UK territorial greenhouse gas emissions statistics 1990-2023" dataset provided by the UK Government. This comprehensive dataset is organized into multiple tables, detailing emissions in million tonnes of carbon dioxide equivalent (MtCO<sub>2</sub>e). The data is broken down by year, economic sector (e.g., electricity supply, transport, industry), greenhouse gas type (e.g., CO<sub>2</sub>, Methane), and fuel source. This rich, time-series data will allow for a multi-faceted analysis of emission patterns.
2. **Data Processing:** The initial data is spread across multiple CSV files corresponding to different tables in the original Excel workbook. The first step will be to consolidate these tables into a structured format suitable for analysis. This will involve merging relevant tables and transforming the data from a wide format (with years as columns) to a long format, which is more appropriate for time-series analysis and modeling. Data cleaning will be performed to handle any missing values or inconsistencies.
3. **Programming Languages and Tools:** The project will primarily be implemented using Python. Key libraries for data manipulation and analysis will include Pandas and NumPy. For data visualization and building an interactive dashboard, I plan to use Matplotlib, Seaborn, and potentially Streamlit. For the modeling phase, Scikit-learn will be used for clustering, and Statsmodels or Prophet will be employed for time-series forecasting.
4. **Data Mining Problem Formulation:** The project will employ a multi-faceted data mining approach. This includes: a **regression** task to forecast future total greenhouse gas emissions using time-series analysis; an **unsupervised clustering** task to group economic sectors based on their historical emission profiles; and comprehensive **exploratory data analysis** to visualize trends and breakdowns by sector and gas, with findings presented in a dashboard. The primary target variable for prediction is total emissions (MtCO<sub>2</sub>e), with year and sector-specific data serving as input features.
5. **Project Type:** This will be an **application-based project**, focusing on applying data mining techniques to a real-world dataset to extract meaningful insights and build predictive models.