Milestone #1 Report

Sai Abhishek Gangineni, Robin Godinho, Marta Gonzalez
INST 737

Introduction

The project idea that our team decided to focus on this semester is about utilizing the data science and machine learning (ML) techniques taught in class to analyze the very crucial issues of climate change and global warming. Beginning with this broad idea, we started our journey by looking into several possible environmental topics of study such as the economy of electric vehicles impacting carbon footprints or the characteristics of safe drinking water to classify bodies of water. Ultimately, it was our search for viable datasets that allowed for us to hone in on a specific question. The ideal dataset that we needed to find had to have at least a thousand data points with multiple features which would allow for prediction through the ML models that we will learn about in class. These characteristics narrowed our search and led us to the dataset that we will cover in the following sections and guided us to create our research question.

Research Question

Predict the amount of Greenhouse Gas (GHG) emissions of a country in one year based on historical data of multiple environmental indicators and features.

Predicting GHG emissions is an important endeavor in terms of both its technical and societal impacts. The predictive ML Models that will be created as a result of this project can serve as a cornerstone for various initiatives aimed at mitigating climate change and adapting to its inevitable impacts.

Firstly, accurate predictions of GHG emissions can facilitate optimal resource allocation. Governments can prioritize investments in renewable energy, energy efficiency, and other low-carbon technologies based on projected emission levels. This would maximize the impact of these investments and accelerate the transition toward a decarbonized economy. It can also support risk assessment and preemptive planning in the face of climate change. By anticipating future emission scenarios, the regions and sectors most vulnerable to climate impacts can be identified beforehand allowing for measures to enhance and protect vulnerable communities and infrastructure from climate-related hazards.

Additionally, by accurately anticipating future emission levels, policymakers, researchers, and organizations can develop targeted measures to curb emissions effectively. With reliable emission predictions, policymakers can design and implement regulations, incentives, and subsidies to target the crucial areas for reducing emissions while encouraging the adoption of cleaner technologies. It can even enable countries to monitor their progress towards their

environmental targets and help fulfill their obligations under international agreements. By honoring their commitments, nations contribute to global efforts to combat climate change and foster international cooperation on environmental issues.

In conclusion, our project to predict greenhouse gas emissions is important because of its potential to advance climate action and sustainable development. From guiding resource allocation and supporting risk assessment to informing policy decisions, the benefits of predictive modeling are numerous.

State of the Art

Article I - Prediction of greenhouse gas emissions for cities and local municipalities monitoring their advances to mitigate and adapt to climate change

This research article is from the *Sustainable Cities and Society* journal, Volume 86. The research in this article proposes a methodology to forecast the GHG emissions of cities monitoring their actions on climate change. The proposed method predicts GHG emissions for cities from their yearly inventories and is illustrated for a group of 1,950 cities and local administrative units in the EU-27. These forecasting models are then validated by the "Leave Last Known Value Out" technique. The idea for this research came about from the Global Covenant of Mayors (GCoM) initiative which involves cities presenting action plans to reduce GHG emissions and adapt to climate change. However, cities often report emissions inventories for arbitrary years, resulting in sparse yearly time series data. So, the paper proposed a machine learning approach to predict emissions for each city's target year, allowing for prediction of emissions while controlling uncertainty and providing reliable information.

Our research question is to predict the GHG emissions for a given country based on historical data on past emission levels. As such, this article is a great source for us to consult as it accomplishes a similar goal, just on a smaller scale of cities and local municipalities. It is therefore imperative for us to include this article in our research to be able to take inspiration from the methodologies utilized to form ML models when creating our own.

Our project aims to both replicate and extend upon the findings of this research paper. We will attempt to emulate the ML predictive model created in this article while extending the scope to accommodate the international scale of our dataset to predict emissions for entire countries.

Article II - Comprehensive green growth indicators across countries and territories

This article is from the *Scientific Data* journal and the goal of this research is to construct green growth measures from 1990 to 2021 across 203 countries. The idea behind this research topic is that a shift towards sustainable green growth is essential for addressing climate change, but the absence of clear definitions and consistent measures poses challenges in identifying its determinants, hindering effective policy making. The metrics identified in the paper are structured around five key dimensions: natural resource base, socio-economic outcomes, environmental productivity, environmental-related policy responses, and quality of life. The researchers utilize a summary index technique that involves a generalized least squares attributed-standardized-weighted index, which addresses issues of correlated variables and missing data. These indicators serve both country-specific and global data modeling purposes, providing valuable insights for policy formulation in the realm of green economic development.

This article was extremely crucial for us when creating our final compiled dataset. Once we had a reliable source of data and a target variable, GHG Emissions, we needed to identify the data features that impact emissions and therefore can help to predict them. This article provided that theoretical context for us by listing out the key environmental indicators which define the "green growth" of a country. With this information, we were able to narrow down on the data tables we wanted to utilize and therefore combine into our final dataset.

By taking the findings of this article as inspiration, our research will endeavor to somewhat replicate and more importantly validate its findings as we attempt to use the same indicators identified in this article to actually create a predictive model for GHG emissions.

Article III - Agriculture-related greenhouse gas emissions and mutation measures

The following research article is an extraction from *Advances in Agronomy* vol.179, chapter 5 p257-376. The research article analyzes the contributions that agriculture, both crop and livestock growth, has had on global greenhouse gas emissions over the years. The article reports that agriculture has The article also assesses a few mitigation measures to increase the food production while curbing the GHG emissions.

The article mentions how the global agriculture and land sectors aim to increase food production by at least 60% between 2010 and 2050 to keep up with the yearly population growth. The goal is to achieve this production rate while limiting the temperature rise below 2 degrees celsius, and reduce GHG emissions by two-thirds.

Our project aims to predict the GHG emissions for a given country based on historical data on past emission levels. It is a well known fact that agriculture is one of the main contributors of high greenhouse gas emissions, and it was imperative to include this article which details the past emissions from the agriculture sector. In the US agriculture has become more efficient, crop and animal production has increased by 30%, while increasing GHG emissions by only 7%.

The goal of our project is to extend upon these findings, and determine which factors are the largest contributors to a country's GHG emissions. Our project will attempt to identify indicators of high GHG emission contributors.

Article IV - G20 Emissions since the Paris Agreement

This research article details the efforts and progress of the G20 group, a group of countries responsible for emitting the majority of greenhouse gas emissions globally. The article describes one of the most recent global efforts to mitigate warming and emissions, the Paris Agreement, and the lack of progress toward lowering global emissions since this policy was created.

The article uses adopted, existing policies, the collection of compiled emissions projections from the Climate Action Tracker, and the comparison to actual historical emissions data to create emissions projections up to 2030. As a result, the research is able to determine an insubstantial change in GHG emissions by 2030 from the G20.

While our research does not pursue policy and its effects on GHG emissions, it is important to note the influence that policy has on global emissions, evaluated in this article. Our project will aim to extend on parts of this article by including the evaluation of historical GHG emissions in our data and further evaluating other related variables. As this article attempts to do, we will be creating our own predictions but through primarily numerical analysis.

Dataset

Data Collection

The first and main part of our dataset that we searched for was the target variable which we chose to focus on, greenhouse gas emissions. Utilizing all the common public access websites like Kaggle, we struggled to find a truly extensive and reliable dataset that wasn't just generated data to be able to use for our research. But, our search bore fruit when we came across the World Resources Institute's Data Lab whose mission is to "enable users to monitor forest change in near-real time, track the drivers of climate change, analyze water risks around the world, examine the intersection of global environmental issues and more." The *Climate Watch* tool for this institute provided the Historical GHG Emissions from 1960 to 2021 for over 195 countries and regions that are United Nations Framework Convention on Climate Change (UNFCCC) member states. The source of this data collected by *Climate Watch* stems from the Paris Agreement which requires participating countries to submit emission inventories that are based on activities within their territory. All the inventories on Climate Watch are based on this production-based accounting.

After finding this dataset with historical data of our target variable, we then set out to search for datasets that cover the data features and environmental indicators which can be linked with GHG emissions thereby being of use in a predictive ML model. Since many of these indicators are also economic indicators, the source for this part of our dataset was found through the World Bank, specifically their World Bank Open Data website which provides free and open access to global development data. Through their search and filter function, we were able to browse by indicator to find relevant data tables that list historical data across similar time scales as the first GHG Emission data for the same number of countries. The indicators that we chose to focus on and collected data for are:

- Electric Power Consumption
- Fossil Fuel Energy Consumption
- Renewable Energy Consumption
- GDP per Capita
- Percentage of Agricultural Land
- Percentage of Forest Area
- Percentage of Urban Population

After collecting and downloading these 9 large datasets as csv files, we then used Excel to go about combining them. First, we copied all of the tables into one Excel sheet with each row representing a single data feature for a particular country and the column representing a year. We kept only the years which had data for all 9 datasets which were historical data from 1990 to 2021. This meant that there were now multiple rows for each country with one for each data feature from the original tables, a total of 18 (Renewable Energy Consumption consisted of several features). So, we created a separate column called *CountryID* which assigned a unique number id to each country. This ultimately created an initial combined dataset with **3021 rows and 36 columns** that consisted of historical data from 194 countries.

Data Cleaning

After the compilation of several datasets was complete, the next step in our process was to complete data cleaning. We chose to use Python, as it was the language we were familiar with, to complete this step. The goal of this step was to upload the dataset as a Pandas dataframe into a Python program before determining the amount of null and missing values in the dataset so that we could determine the appropriate method of handling them, whether it is to remove or disregard. The steps we followed include:

Step 1: Understanding the dataset

- The original compiled dataset consisted of 3021 rows x 36 columns.

Step 2: Determine if the dataset consisted of any null values

- There were several missing data points.

Step 3: Count the number of data points for each column

- It was determined that there were several columns with null attributes particularly from the years 1990-1999 and 2016-2021.

Step 4: Remove columns and rows missing a significant number of data points

- Use the 'iloc' method to remove columns missing significant data points.
- Remove rows missing significant data points using dropna() function from the pandas module.

After completing this Data Cleaning procedure, our Final Dataset now consisted of **2664 rows and 19 columns** with historical data for **194 countries** from **2000 to 2015.**

Final Dataset Description

The columns of the dataset begin with *CountryID* and *Country Name* followed by the *Data Feature* and finally the year columns from 2000-2015. The number of rows, meaning data features, for each country consisted of 18 features which are:

- GHG Emissions (MtCO2e)
- Electric power consumption (kWh per capita)
- Fossil fuel energy consumption (% of total)
- GDP per capita, PPP (current international $)
- Agricultural land (% of land area)
- Forest area (% of land area)
- Urban population (% of total population)
- Access to Clean Fuels and Technologies for cooking (% of total population)
- Access to electricity (% of rural population with access)
- Access to electricity (% of total population)
- Access to electricity (% of urban population with access)
- Energy intensity level of primary energy (MJ/2011 USD PPP)
- Renewable electricity output (GWh)
- Renewable electricity share of total electricity output (%)
- Renewable energy consumption (TJ)
- Renewable energy share of TFEC (%)
- Total electricity output (GWh)
- Total final energy consumption (TFEC) (TJ)

Data Exploration

We began our data exploration efforts by evaluating the structure of our dataset. As a result of our data features being values in a column versus columns themselves, we decided to create subsets of the data based on our individual data features. It was important for us to get an initial understanding of our features so we conducted global summary statistics based on year. This was useful for providing us with a baseline for what the distribution for each year normally looked like and allowed us to identify any years that deviated drastically from the typical distribution.

The next step was analyzing the distribution of each feature by country in order to determine whether there were any outliers in the data. To do this, we began by melting our subsets so that we could visualize them. We also split each subset into additional subsets separated by the first letter of each country name. We did this in order to better visualize the data and prevent clumping and a lack of clarity in our visualizations. Once this was complete, we created boxplots to visualize the distribution of each country by feature and identified all countries with outliers for each feature.

Our last step was to visualize the relationship between each variable and GHG emissions. We completed this by creating scatter plots for every variable relationship against GHG emissions. This allowed us to get an initial grasp of the relationships as they pertained to GHG emissions and whether they initially seemed to be strong or weak relationships.

Contributions

Sai Abhishek Gangineni - 40% of Report, 20% of Code, 33.33% of Presentation

Robin Godinho - 30% of Report, 40% of Code, 33.33% of Presentation

Marta Gonzalez - 30% of Report, 40% of Code, 33.33% of Presentation

References

State of the Art:

Franco, C., Melica, G., Treville, A., Baldi, M. G., Pisoni, E., Bertoldi, P., & Thiel, C. (2022). Prediction of greenhouse gas emissions for cities and local municipalities monitoring their advances to mitigate and adapt to climate change. Sustainable Cities and Society, 86, 104114. https://doi.org/10.1016/j.scs.2022.104114

Sarkodie, S.A., Owusu, P.A. & Taden, J. Comprehensive green growth indicators across countries and territories. Sci Data 10, 413 (2023). https://doi.org/10.1038/s41597-023-02319-4

Nascimento, Leonardo, et al. "The G20 emission projections to 2030 improved since the Paris Agreement, but only slightly." National Library of Medicine, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9281192/. Accessed 25 02 24.

Pasricha, N. S., Ghosh, P. K., & Singh, R. (2023). Agriculture-related greenhouse gas emissions and mitigation measures. In D. L. Sparks (Ed.), Advances in Agronomy (Vol. 179, pp. 257-376). Academic Press. https://doi.org/10.1016/bs.agron.2023.01.005

GHG Emissions Dataset:

https://www.wri.org/data
https://www.climatewatchdata.org/

Global Carbon Project. (2023). Supplemental data of Global Carbon Budget 2023 (Version 1.1) Data set. Global Carbon Project. https://doi.org/10.18160/gcp-2023

Environmental Indicator Datasets:

https://data.worldbank.org/