Mid-Point Deliverables - Climate TRACE Team

1. Main Deliverable: EUI Estimation Technique
The main deliverable for this milestone is the Energy Use Intensity (EUI) estimation technique. This technique leverages globally available features that enable EUI prediction. By selecting these key features, we aim to generate the first iteration of EUI predictions. The goal for this semester is to achieve a Mean Absolute Percentage Error (MAPE) in the range of 30-40%, which we consider an ideal target. However, it is possible that we may not reach this range at this stage, and we will focus on refining and improving this technique in the following semester, with the final goal of enabling EUI prediction anywhere in the world.

1.1 Features and Pipeline
This section corresponds to the selection of features for the EUI estimation model. It explains the reason for selecting these features and highlights their relevance. Additionally, it includes the data pipeline for processing and cleaning these features to make them ready for modeling. The features we are considering and will be testing in the models are

- Heating Degree Days (HDD)
- Cooling Degree Days (CDD)
- Temperature
- Humidity
- Latitude
- Longitud
- GDP per capita
- Urbanization Rate
- Paris agreement participation
- Human Development Index (HDI)
- Educational index
- Income index

It's worth mentioning that for variables like temperature, humidity, and HDD, since we have detailed data both geographically and temporally, we can explore different approaches to transforming the variable. For example, we could use an annual average for the closest point or consider an average within a surrounding radius.

Format:
The deliverable will include a description of the features, their sources, and the reason for their selection. This information will be included in the README file. Additionally, a Python-based pipeline will be delivered in Jupyter Notebook format, which processes the initial input data, performs data cleaning, and handles tasks like merging to ensure the data is ready for modeling.

1.2 Model Development and Experimentation
This section outlines the process of testing different models and experimental designs, including the exploration of the features mentioned earlier, to identify the best approach for EUI prediction. We will explore different machine learning models, including:

- Linear Regression: Baseline model to establish an initial reference.
- Random Forest
- K-Nearest Neighbors (KNN)
- Ensemble models: XGBoost and CatBoost

Given the challenge of regional variations in global data, we will validate our predictions at the regional level across 5 regions using 3 strategies to identify biases and improve model robustness.

- Within-Domain: We train our model on 80% of the data from each region and test it on the remaining 20% of the same region. To obtain a global result, we calculate the average of our evaluation metrics across the 5 regions.

- Cross-Domain: We train our model using 80% of the data from 4 regions and test it on the 20% of the 5th region. We repeat this process to evaluate the model's extrapolation. The global result is the average of all outcomes.

- All-Domain: We use 80% of the data from all 5 regions for training and test it on the 20% of the first region. This process repeats for each region. The global result is the average across all tests.

We aim to assess our model's generalization by comparing its performance within the same region (Within-Domain) and its ability to extrapolate to other regions (Cross-Domain). The goal is to reduce the gap between these strategies to improve accuracy and understand extrapolation errors. Additionally, we want to understand if there are regions that perform better than others in specific outcomes, which can help us tailor our model to regional differences.

For this evaluation, we will use the following metrics:

- Mean Absolute Percentage Error (MAPE), with a target range of 30-40%. While this is our primary target, it is also important to consider other metrics to assess model performance comprehensively.
- R^2
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Weighted Absolute Percentage Error (WAPE)

And also for diagnose how well our model is performing, we will generate the following plots:

- Actual EUI vs. Predicted EUI
- Error Distribution Plot (Residuals)

Format:
This analysis will be developed in a Jupyter Notebook, presented in a slide deck, and key conclusions will be summarized in the README file.

1.3 Analysis and Conclusion

This section will summarize the results from the experimental design and evaluation, providing an analysis of the model's performance. It will also offer recommendations for refining the model and suggestions for additional features or data sources that could improve predictions in the next semester.

Format:

The deliverable will include a written summary of findings and next steps, which will be detailed in a slide deck and summarized in the README file.