# EN 605.662 Data Visualization

## Final Project: Climate Change

**Student Name: Raouf Eskandar**

**Dataset: realistic\_climate\_change\_impacts.csv**

**Dataset Source: https://www.kaggle.com/datasets/bhadramohit/climate-change-dataset**

**Abstract**

This project explores the tangible effects of climate change by analyzing country-level data on extreme weather events, CO₂ emissions, temperature anomalies, economic losses, and population displacement between 2000 and 2023. Using the dataset sourced from [Kaggle](https://www.kaggle.com/datasets/bhadramohit/climate-change-dataset), we applied machine learning techniques—such as clustering and classification—and created an interactive dashboard using Streamlit. A novel element of this project is the integration of a GPT-2 model to generate country-specific insights and simulate interactive, AI-powered analysis. The goal was to turn raw data into meaningful, interpretable, and engaging visual insights.

**Introduction**

Climate change continues to accelerate, manifesting through rising global temperatures, increased CO₂ emissions, and a growing frequency of natural disasters. While scientists have long documented these effects, making sense of the data and presenting it in a user-friendly way remains a challenge. This project aims to bridge that gap by offering an intuitive interface that not only visualizes climate trends but also provides AI-driven summaries to help users understand regional impacts.

**Background**

Numerous studies have examined the impact of greenhouse gases, rising temperatures, and natural disasters. Many of these rely heavily on statistical models and static charts. However, there's a growing demand for tools that are both interactive and insightful. Recent developments in AI—especially language models like GPT-2—offer new ways to interpret and communicate findings. By combining machine learning, data visualization, and NLP models, we aim to push the boundary of how climate data is explored. Recent advancements in Natural Language Processing, particularly the GPT-2 model by Radford et al. [2] and the Hugging Face Transformers library [3][4], enable the automated generation of human-like summaries. These models have proven effective for contextual text generation [5], making them ideal for AI-driven interpretations of climate data.

Effective climate communication has been a growing concern in visualization research. Doudkin [13] emphasized the importance of using interdisciplinary visualization methods to make climate data accessible to non-expert audiences. This aligns with our use of interactive dashboards and AI-generated summaries to improve comprehension.

**Approach**

**Dataset Overview**

The dataset used is titled *realistic\_climate\_change\_impacts.csv*, downloaded from Kaggle. It includes over 10,000 records spanning 2000–2023 and contains the following fields:

* Country: Affected nation
* CO₂ Level (ppm): Atmospheric carbon dioxide concentration
* Temperature Anomaly (°C): Deviation from the baseline temperature
* Economic Impact (USD): Estimated monetary damage
* Population Affected: Number of people impacted
* Extreme Weather Event: Type of event (e.g., flood, drought, wildfire)

Missing values were addressed using imputation. For example, blank *extremeweatherevent* fields were filled with “Unknown” to preserve record integrity.

The use of machine learning and AI-based text generation is consistent with the broader movement advocated by Climate Change AI [14], which encourages leveraging modern ML techniques to support mitigation, adaptation, and communication around climate change.

**Technologies Used**

We structured our work into two main parts: analysis and visualization.

**1. Jupyter Notebook (EDA + ML)**

* Data Cleaning: Removal of anomalies, conversion of strings to numeric, handling null values.
* K-Means Clustering: Grouped events based on shared features like CO₂ levels and economic loss. Optimal clusters were found using the elbow method (k=4).
* Random Forest Classification: Trained a classifier to predict event type. Despite a lower accuracy (13.8%), it provided valuable insights on feature importance.
* Principal Component Analysis (PCA): Reduced data dimensionality to better visualize clustering.

**2. Streamlit Dashboard**

Streamlit was used to turn the analysis into an interactive dashboard with the following features:

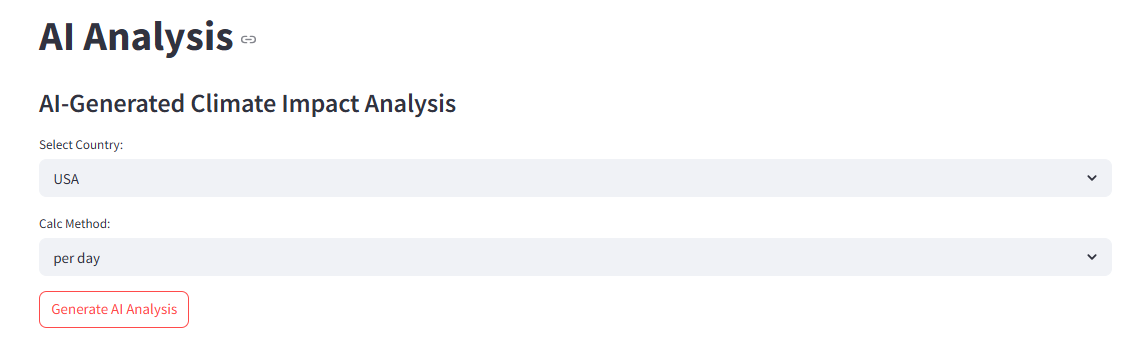
* Page navigation for different visualizations
* Data filtering via sliders (e.g., CO₂ range, temperature anomaly)
* Interactive maps, word clouds, and model outputs

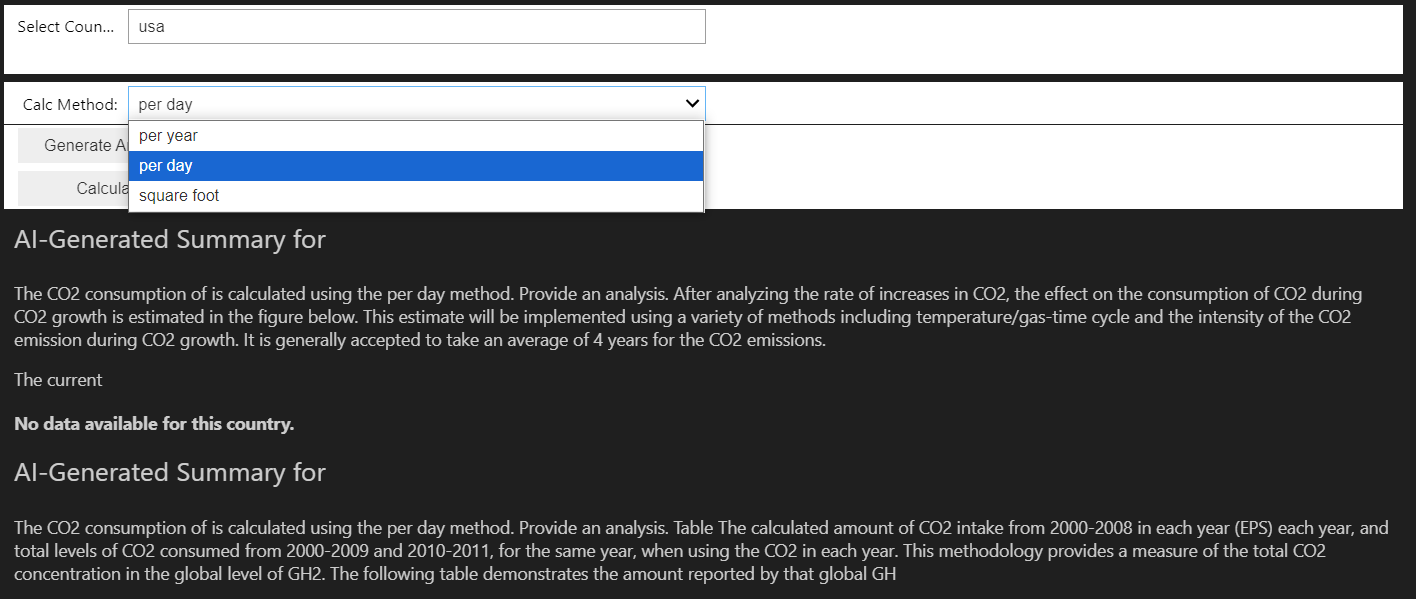
**3. AI Integration Using GPT-2**

We leveraged Hugging Face’s Transformers library to integrate GPT-2 into the dashboard. Here’s how it works:

* Users select a country
* Average CO₂ levels are retrieved
* A prompt is dynamically constructed and sent to the GPT-2 model
* The model generates a short summary of the country’s climate situation
* Based on user input (e.g., “per day”), a corresponding economic impact is calculated

All analysis was conducted using Python-based tools, including pandas [6] for data wrangling, scikit-learn [7] for machine learning models, matplotlib [9] and Plotly [10] for visualizations, and Streamlit [12] for an intuitive dashboard interface.

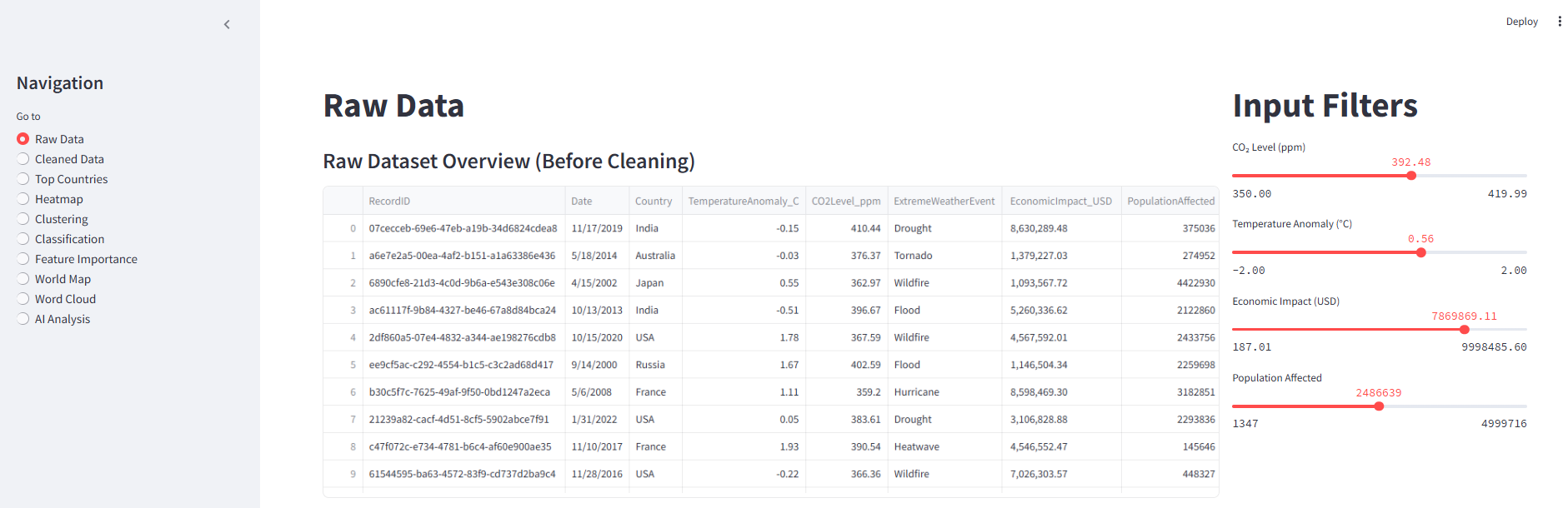




**Results**

**Streamlit Dashboard Structure**

The dashboard consists of three main sections:



**A) Navigation Panel**

Located on the left, it lets users switch between:

* Raw Data: View unprocessed dataset
* Cleaned Data: Post-EDA data
* Top Countries: Most affected nations
* Heatmap: Correlation matrix
* Clustering: KMeans visual output
* Classification: RF prediction and confusion matrix
* Feature Importance: RF feature weighting
* World Map: Geographic event count
* Word Cloud: Common event descriptors
* AI Analysis: GPT-generated summaries + economic impact calculator

**B) Shown Data**

Displays data tables, plots, or AI output depending on the selected tab. Visuals are rendered using Plotly, Matplotlib, or Seaborn.

**C) Filter Section**

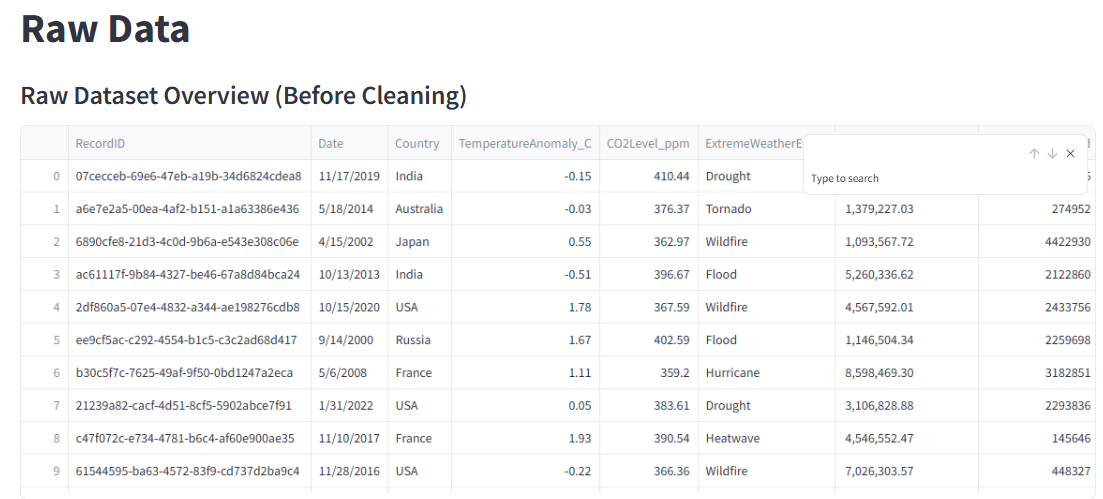
Adjust input filters to refine analysis:

* CO₂ Level Slider
* Temperature Anomaly Slider
* Economic Impact Slider
* Population Affected Slider

Each slider dynamically updates results shown in the visual panel.

**Dashboard of the Navigation Panel (A)**

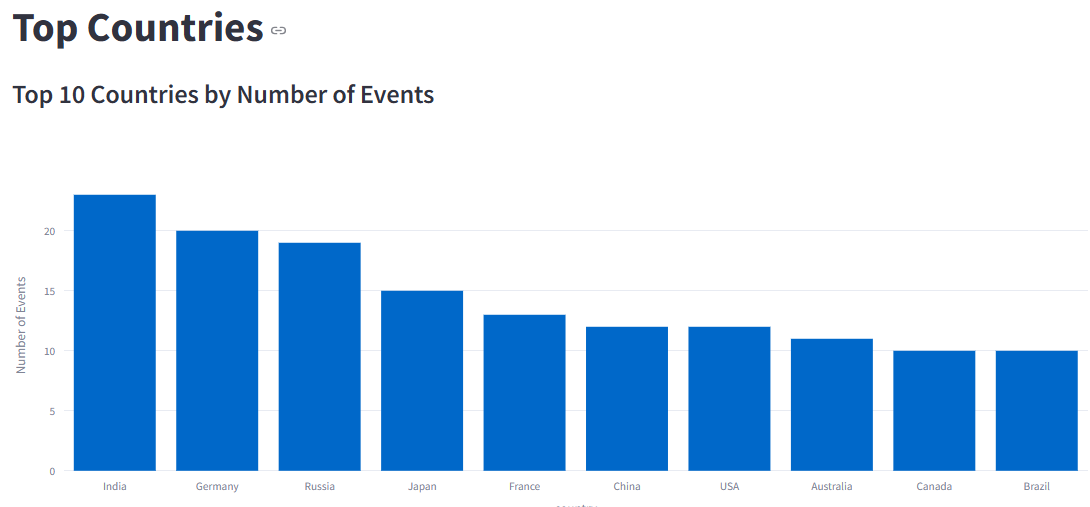
* Raw Data: View unprocessed dataset



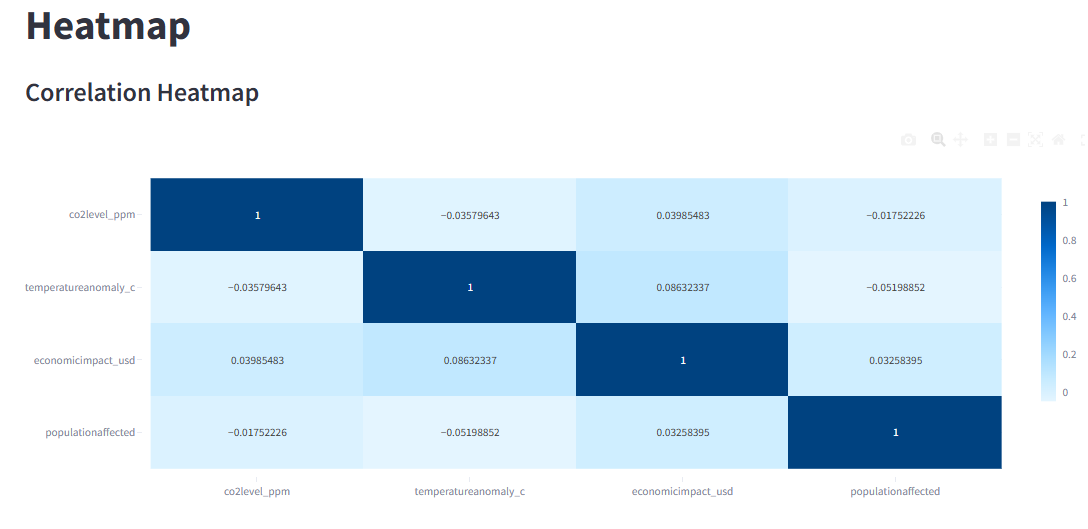
* Cleaned Data: Post-EDA data



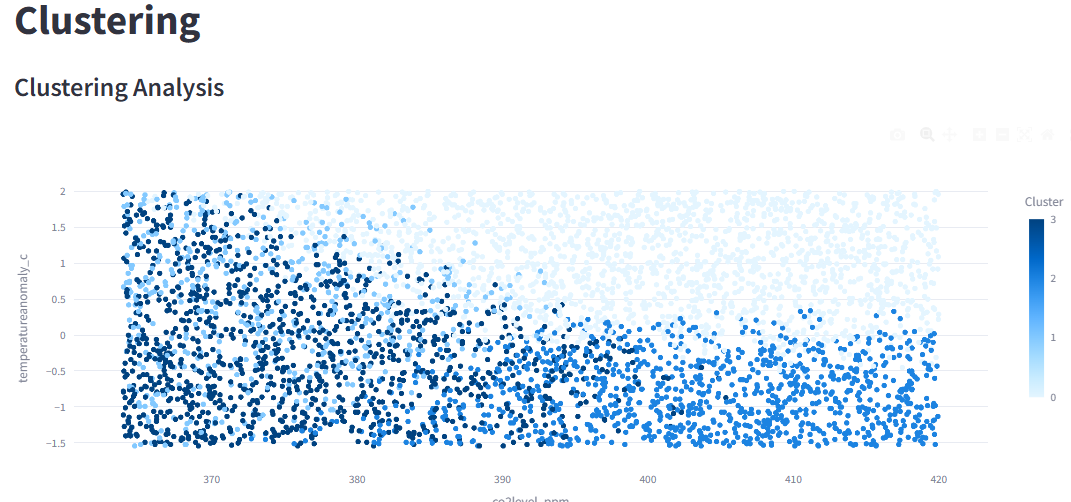
* Top Countries: Most affected nations



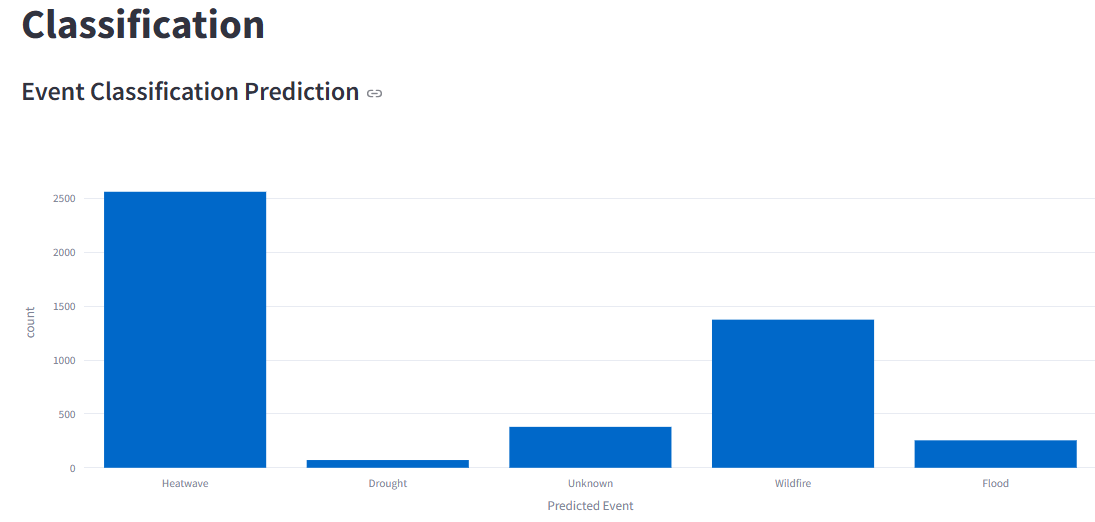
* Heatmap: Correlation matrix



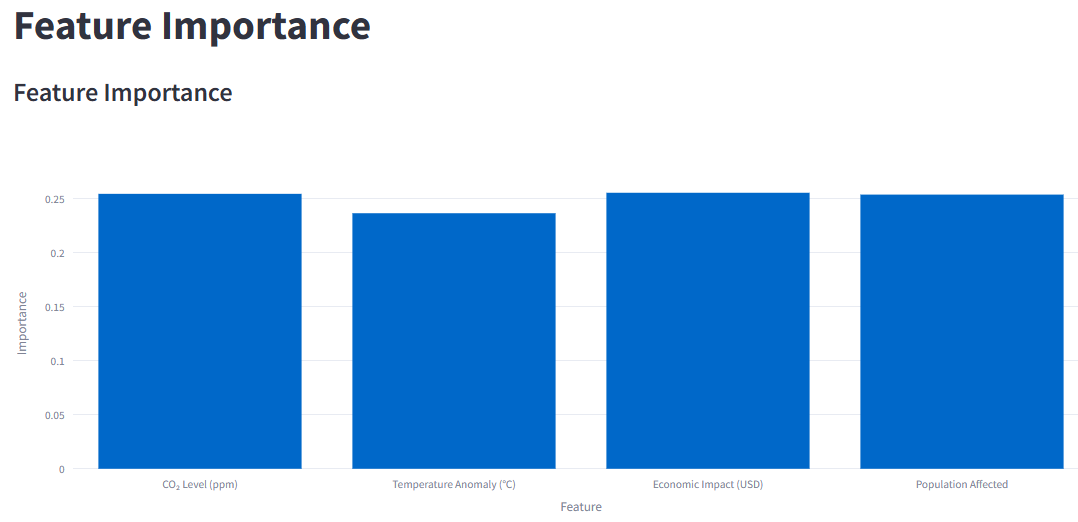
* Clustering: KMeans visual output



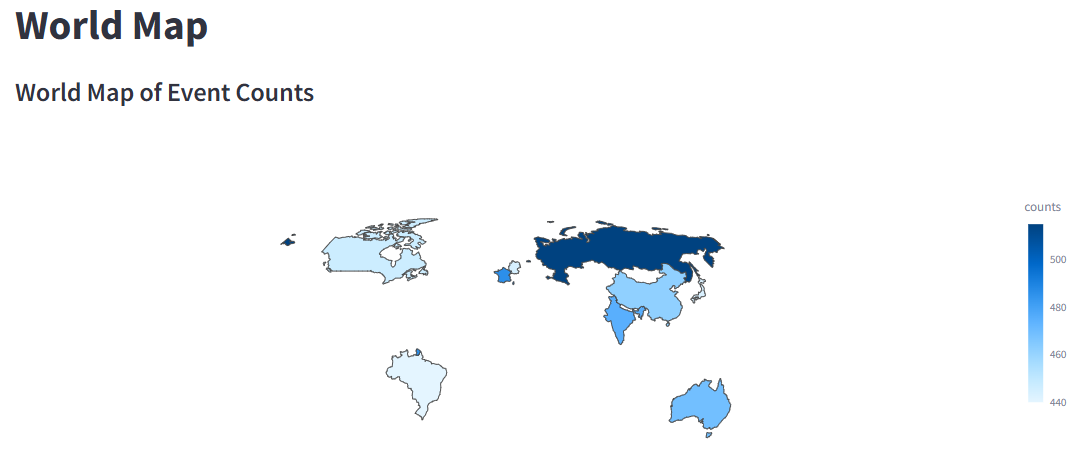
* Classification: RF prediction and confusion matrix



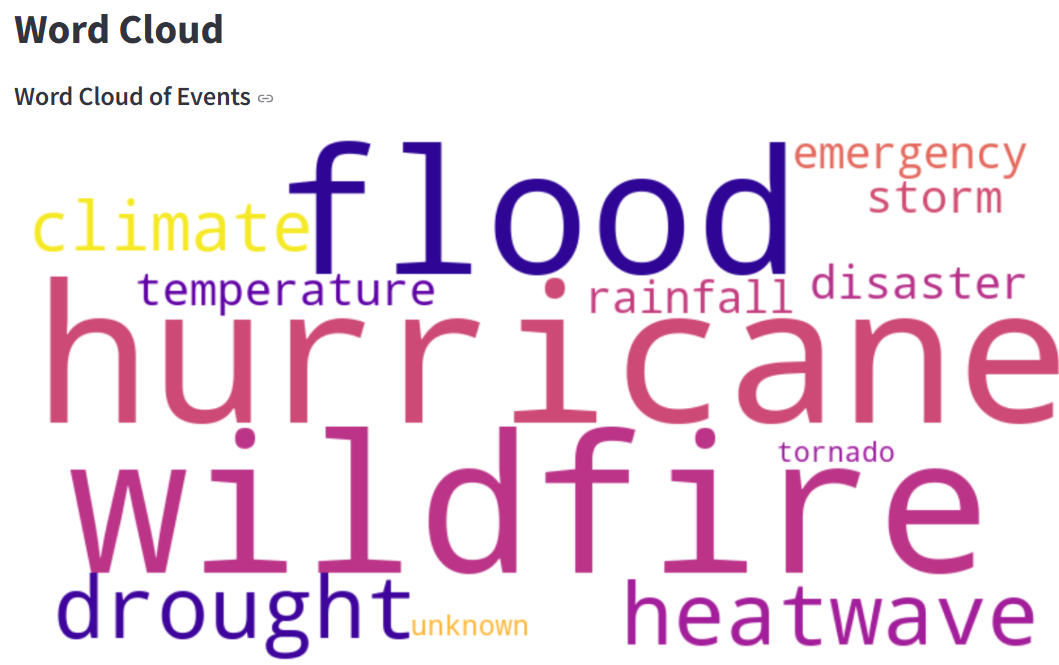
* Feature Importance: RF feature weighting



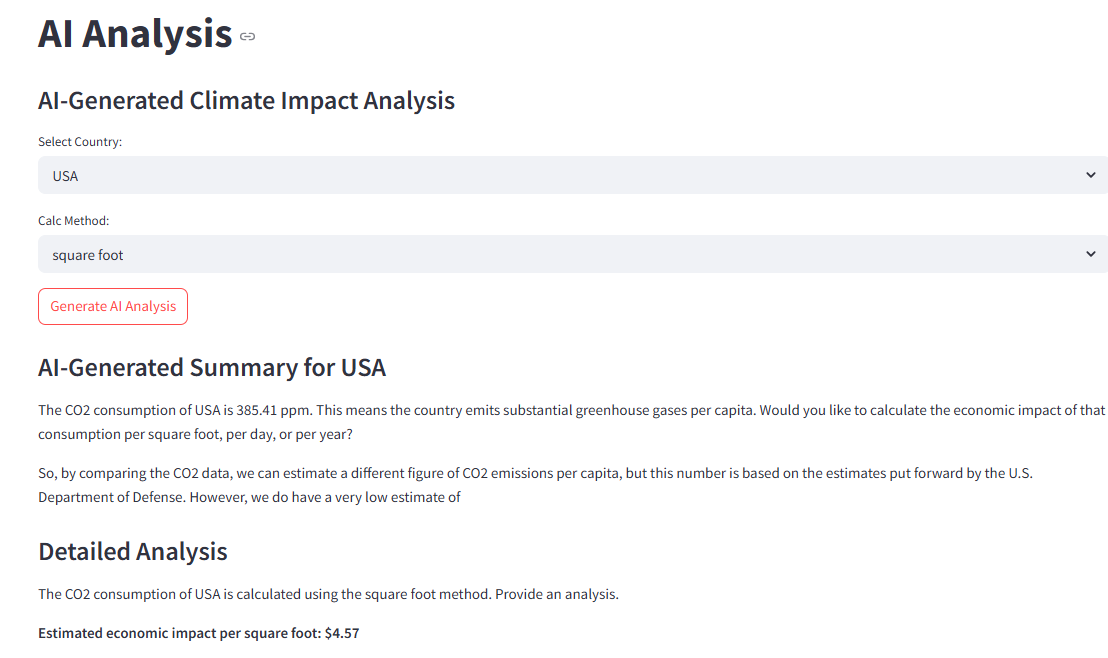
* World Map: Geographic event count



* Word Cloud: Common event descriptors



* AI Analysis: GPT-generated summaries + economic impact calculator . [2], [3], [4], [5] — these support the choice of GPT-2 and the method of AI integration



**Conclusion**

This project blends data science, machine learning, and NLP to deliver a comprehensive and accessible overview of climate change impacts. Although our Random Forest model yielded modest accuracy, it highlighted areas for improvement—such as expanding the dataset or engineering new features. More importantly, the integration of GPT-2 demonstrated how AI can make climate data more understandable for broader audiences. The Streamlit dashboard provides a valuable tool for researchers, educators, and policymakers alike. While conventional methods offer strong numerical insights, the addition of transformer-based NLP models [4] demonstrates how AI can democratize data access and improve decision-making in climate policy.

**References**

[1] Bhadra, M. (2023). *Climate Change Dataset*. Kaggle. <https://www.kaggle.com/datasets/bhadramohit/climate-change-dataset>

[2] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2019). *Improving language understanding by generative pre-training*. OpenAI. <https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf>

[3] Hugging Face. (n.d.). *Transformers Documentation*. Hugging Face. <https://huggingface.co/docs/transformers/index>

[4] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). *Transformers: State-of-the-art natural language processing*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. <https://www.aclweb.org/anthology/2020.emnlp-demos.6>

[5] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language models are few-shot learners*. arXiv. <https://arxiv.org/abs/2005.14165>

[6] McKinney, W. (2010). *Data structures for statistical computing in Python*. Proceedings of the 9th Python in Science Conference, 51–56. <https://doi.org/10.25080/Majora-92bf1922-00a>

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>

[8] van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). *The NumPy array: A structure for efficient numerical computation*. Computing in Science & Engineering, 13(2), 22–30. <https://doi.org/10.1109/MCSE.2011.37>

[9] Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>

[10] Plotly Technologies Inc. (2015). *Collaborative data science*. Montréal, QC: Plotly Technologies Inc. <https://plotly.com/python/>

[11] Waskom, M. (2021). *Seaborn: Statistical data visualization*. Journal of Open Source Software, 6(60), 3021. <https://doi.org/10.21105/joss.03021>

[12] Streamlit Inc. (n.d.). *Streamlit: The fastest way to build data apps*. <https://streamlit.io/>

**[13]** Doudkin, A. (2024). *Interdisciplinary methods of digital climate change data visualization for a non-expert audience*. arXiv. <https://arxiv.org/abs/2401.03532>

**[14]** Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., ... & Bengio, Y. (2019). *Tackling Climate Change with Machine Learning*. Climate Change AI. <https://www.climatechange.ai/papers>