INTRODUCTION TO DATA SCIENCE
MILESTONE - III
REPORT

TABLE OF CONTENTS

Project Title: Predictive Analytics for Climate Variables and Disaster Severity Using Supervised and Unsupervised Learning

Author : Nikhitha Nagalla

UFID : 66288276

1. INTRODUCTION

1.1 OBJECTIVE

This project aims to develop an advanced recommender system that predicts climate temperature trends and assesses potential natural disasters based on the user-specified year and location (country). By leveraging historical data and predictive modeling, the system will provide valuable insights into environmental risks, enabling users to make informed decisions. This tool will be a crucial resource for climate analysis, disaster preparedness, and risk mitigation strategies.

1.2. TYPE OF TOOL

This project will develop a **Climate and Disaster Prediction Recommender System**, which will

- Recommends expected **temperature variations** according to time and location.
- Predicts the likelihood of **natural disasters** based on past trends.
- Deliver insights through an **interactive interface**.

1.3. DATA SOURCES

1. Global Temperature Records(1850-2013) - kaggle
2. Per capita Co2 emissions - Ourworldindata
3. Annual precipitation, 1940 to 2024 - Ourworldindata
4. Deforestation and Forest Loss - Ourworldindata
5. Natural Disasters Emergency Events Database & Country Profiles - Omdena

1.4. TECHNOLOGY STACK

Programming Language: Python

Libraries and Frameworks

- **Data Manipulation:** Pandas, NumPy
- **Visualization:** Matplotlib, Seaborn, Plotly
- **Machine Learning & Prediction:** Scikit-Learn, TensorFlow
- **Geospatial Analysis:** Geopandas, Folium
- **Report Generation:** Google docs

2. PROJECT TIMELINE

s.no	Task	Deadline
1.	Data collection and preprocessing	23rd February 2025
2.	Exploratory Data Analysis	23rd February 2025
3.	Feature Engineering and Feature selection	11th March 2025
4.	Data Modeling	21st March 2025
5.	Evaluation and testing	15th April 2025
6.	Interpreting and visualising results	23rd April 2025

3. SUMMARY OF MILESTONE 1 - EDA

Global Temperatures Are Rising – There is a consistent increase in global temperatures over time, with certain regions experiencing extreme heat anomalies and fluctuations.

CO_2 Emissions Directly Correlate with Climate Change – Countries with high industrial activity show higher temperature increases, confirming the impact of human-induced emissions on global warming.

Natural Disasters Are Becoming More Frequent & Severe – The number of climate-related disasters (floods, hurricanes, wildfires, etc.) has significantly increased, leading to higher economic losses and casualties.

Some Regions Are at Higher Risk – Developing countries and coastal regions face greater disaster risks due to higher exposure, weaker infrastructure, and poor disaster preparedness.

Predictions Can Aid Climate Action – Data-driven forecasting can help governments and organizations prepare for future climate risks, improve disaster response strategies, and implement sustainable policies.

The five datasets were consolidated into one unified dataset, "merged_df.csv," which served as the foundation for both classification and regression analyses.

Handling Missing Values

The columns Code", 'AverageTemperatureUncertainty', 'City', 'ISO', 'Disaster Group', 'Disaster Subroup' are removed using drop() method.

Some columns like co2 emission, average temperature, Precipitation columns are filled with the average value within the particular columns.

The following are some of the visualizations to understand the data better.

This diagram is a grid of distribution plots (histograms with KDE curves) that visually represents the distribution of values for multiple numerical features in your dataset.

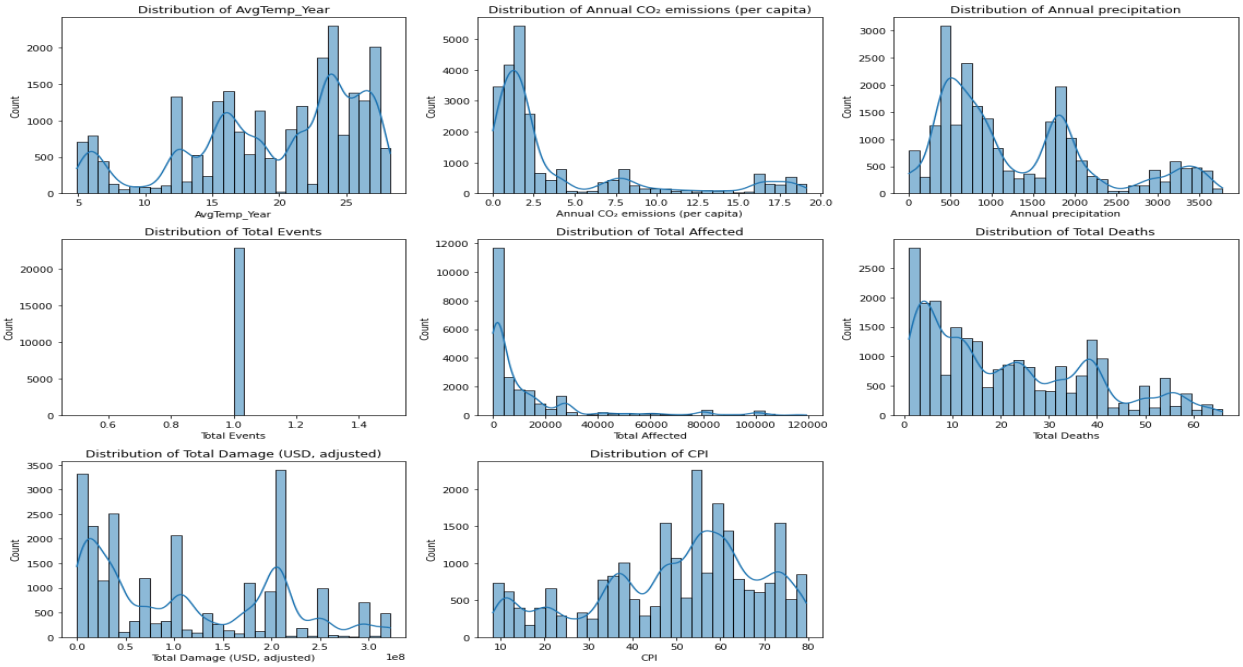


Fig.1. Histograms of different features

Skewed Distributions: Several plots (Total Affected, Total Damage, etc.) are right-skewed, meaning in most instances the values are low, a few are of extremely high magnitude (extreme disasters). Uniform/Normal-like: Features such as AvgTemp_Year and Year appear to be evenly distributed, implying they are covering a wider and more generalized range of data. Zero-heavy columns: Certain columns (maybe Mass movement (dry) or Volcanic activity) have very high spikes at zero, indicating that perhaps various countries do not have these disasters.

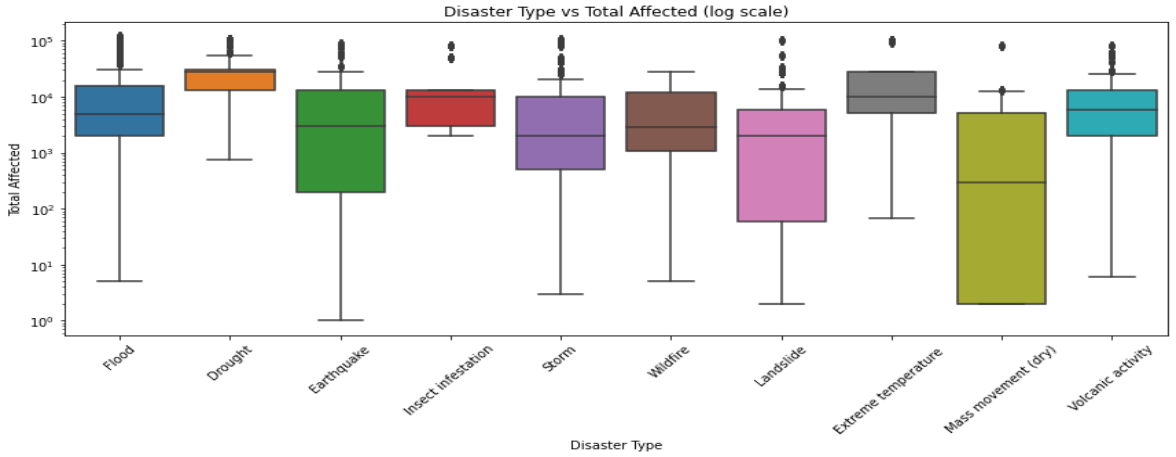


Fig.2. Boxplot regarding the disasters

Most features appear to have outliers in the boxplot which is mostly due to extreme disaster events. Some of the features such as Annual precipitation have wider IQRs suggesting much variability, while some features show an even spread. It also contains features with skewed distributions, with right-skewness in features where the median lies near the bottom of the box.

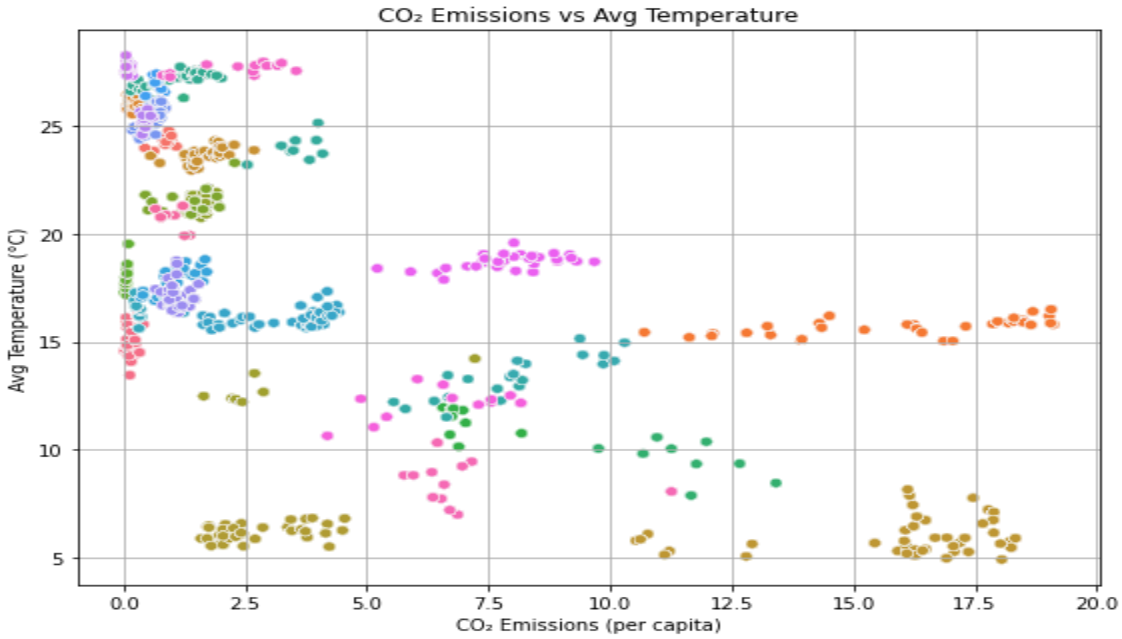


Fig.3. Average Temperature vs CO2 emissions

The scatter plot above reveals clusters of countries based on PCA components. The individual points represent countries, while the colors denote different clusters formed by a clustering algorithm. In an overall spatial grouping, similar disaster profiles, for example, flood-prone or drought-prone, create contours of these clusters, which are sharply defined and distinct from each other.

4. SUMMARY OF MILESTONE 2 - MODEL SELECTION AND EVALUATION

4.1. Feature Engineering

A new categorical target, Disaster Impact Severity, was created by binning Total Affected into Low, Medium, and High impact levels. Average Yearly Temperature (AvgTemp_Year), calculated from monthly records, makes the climate more comprehensible. Dictionaries of emissions, precipitation, and disaster history were merged, giving each country a holistic environmental profile for a year. Categorical features were encoded as Country & Disaster Type, using label encoders. Latitude and longitude are converted into a float format string for geospatial modeling.

```python
def categorize_impact(x):
    if x < 1000:
        return "Low"
    elif x <= 50000:
        return "Medium"
    else:
        return "High"

merged_df["Impact_Level"] = merged_df["Total Affected"].apply(categorize_impact)
```

Fig.4. Code snippet for new_feature “Impact Level”

4.2 Feature Selection Techniques

Correlation Heatmap eliminated the multicollinearity, where AvgTemp_Year, CO_2 emissions, and Total Damage remained as informative features, while redundant features were discarded. Random Forest Feature Importance further highlighted a few key drivers in the classification process, thus making it easier to understand disaster impact more by showing the contribution of mostly climate and socio-economic attributes. However, PCA (Principal Component Analysis) discovered that the first few components could hold over 90% of the variance brought to dimensionality reduction, where it is necessary.

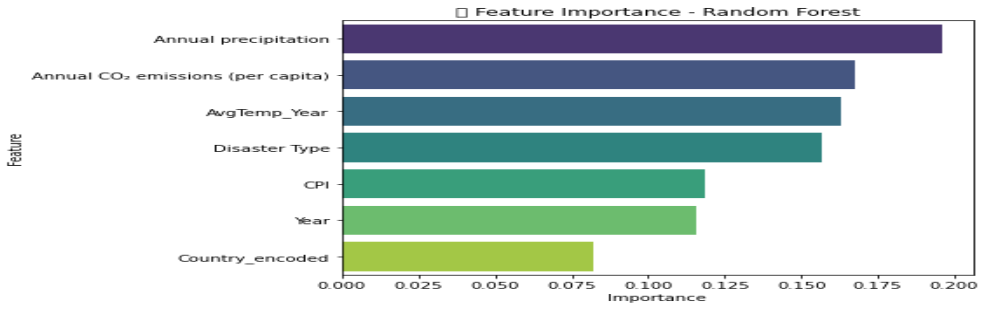


Fig.5. Feature Importance - Random Forest

4.3. Principal Component Analysis (PCA)

To improve the model performance and reduce possible correlation among features, the dimensionality reduction technique of Principal Component Analysis (PCA) was utilized. More than 90% of the total variance in the data is explained by the first four principal components. The curve starts to flatten after the 6th component, indicating the diminishing significance of additional components.

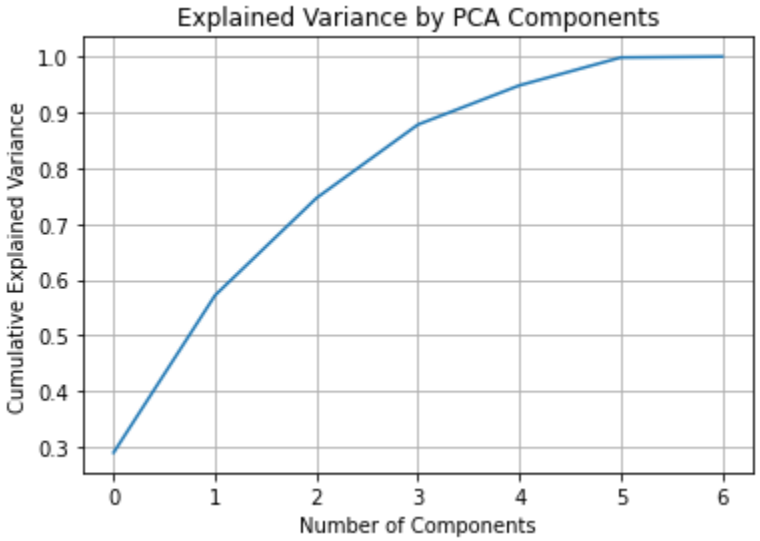


Fig.6. PCA explaining 90% variance

4.4 Model Training

In this phase, both classification and regression tasks are performed.

- A classification task for predicting the severity of disaster impact and the type of disaster for a given country-year-disaster combination.
- A Regression task for predicting future climate indicators such as average annual temperature, CO_2 emissions, and precipitation for any given country and year.
  An outline of the Data Modeling
  To validate the generalization of models, the data was split into 80% of the training sets and 20% for the test set. Two main tasks were performed:

Classification Task - Predicting disaster effect severity
Models used: Random Forest, MLP (neural network), and XGBoost

Performance:

- The best result, around 97% accurate, highly precise/recalled with minimized classification errors, was achieved by XGBoost.
    -
- It adeptly differentiated among Low, Medium, and High impact categories.

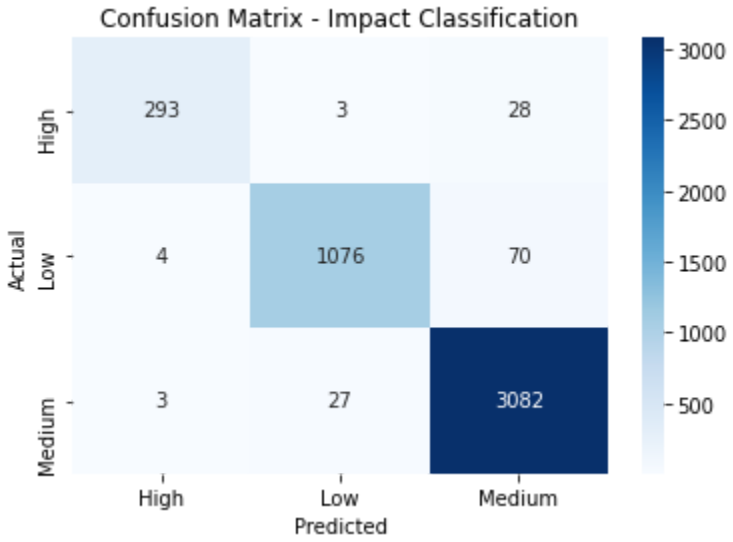


Fig.7. Confusion Matrix - XGBoost

Regression Task – Predict Climate Indicators with the target variables Temperature, CO_2, and Precipitation
Models used: Random Forest Regressor, LSTM, XGBoost

Performance

Regression task to predict climate indicators with targets being temperature, CO_2 emissions, and precipitation. Models were: random forest regressor, LSTM, and XGBoost. The most significant score was nearly perfect in all cases and had the lowest MAE/RMSE across all targets.
Random Forest and LSTM also presented fruitful results, but did not match this overall precision.

```
142/142 [==============================] - 0s 2ms/step

🔍 Temperature Evaluation:
  RMSE     : 0.0949
  R² Score : 0.8791

🔍 CO₂ Emissions Evaluation:
  RMSE     : 0.0673
  R² Score : 0.9344

🔍 Precipitation Evaluation:
  RMSE     : 0.0862
  R² Score : 0.8844
```

Fig.8. Evaluation Metrics - LSTM

```
1/1 [==============================] - 0s 147ms/step

{'AvgTemp_Year': 13.127828367014725,
 'Annual CO₂ emissions (per capita)': 6.227190298558802,
 'Annual precipitation': 1753.4327202441154}
```

Fig.9. Forecast Results - LSTM

Insights

- For classification, XGBoost outperformed other models with the highest accuracy and was particularly effective in predicting medium-impact disasters, which form the majority class.
- Confusion matrices showed minimal misclassification between Low and High impact classes, indicating strong decision boundaries.
- For regression, LSTM performed well, especially for temperature and CO_2, but had slightly less precision than Random Forest and XGBoost.

- LSTM is effective with temporal sequences. The feature set (Country + Year) was highly predictive of climate trends, especially when the model learned from rich historical data.

5. CLUSTERING

Unsupervised clustering techniques used to derive latent structures and country groupings with respect to exposure and environmental characteristics were configured to cut countries into meaningful clusters for comparative risk analysis and for the development of policy.

Feature Preparation:

Disaster Frequency Normalization: Normalization of different disaster types (Floods, Earthquakes, Storms, and Wildfires) by population and area to handle scale differences across societies.

Dimensionality Reduction: Principal Component Analysis (PCA) was applied to reduce the dimensionality of the normalized feature space for better visualization and reduced redundancy.

Algorithms Used:

K-Means Clustering

Clustering is performed using five clusters. Cluster labels were interpreted based on dominant disaster types and exposure levels.

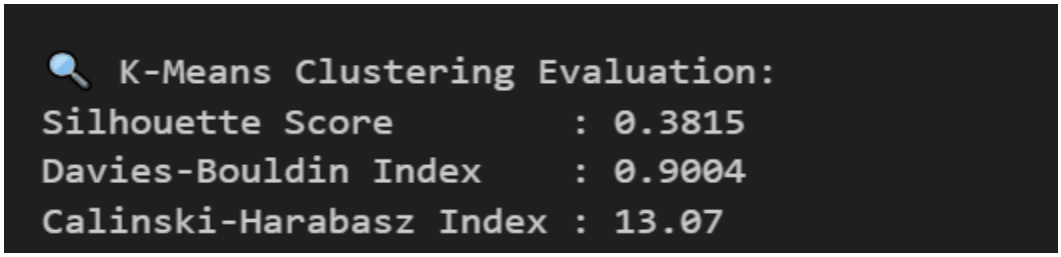


Fig.10. Regression Metrics - KMeans

Agglomerative Clustering

Clustering is performed with 5 clusters, similar in trend to K-Means but more compact. Cluster labels were manually interpreted and assigned based on dominant normalized disaster profiles. Choropleth maps and PCA projections provided clear geographic groupings. Each method provided different cluster boundaries and distributions, offering diverse insights into country groupings. It has a Silhouette Score of around 0.35, Davies-Bouldin Index of around 0.89 and Calinski-Harabasz Score of 12.64

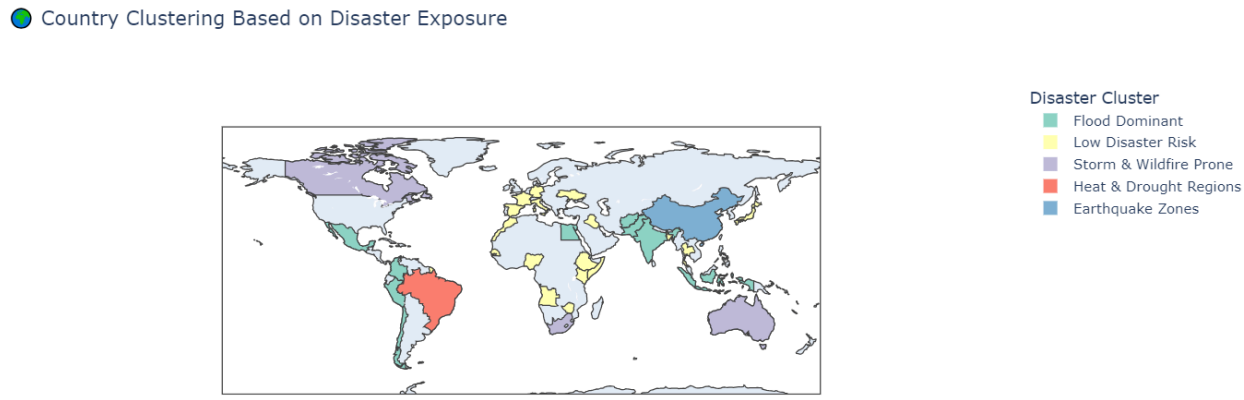


Fig.11. Clustered Countries - Agglomerative clustering

GMM Clustering

It provided probabilistic cluster memberships and captured overlapping boundaries between similar disaster zones. It is slightly more flexible than hard-assignment models, but with a more complex interpretation. Cluster labels are interpretations based on their dominant features.

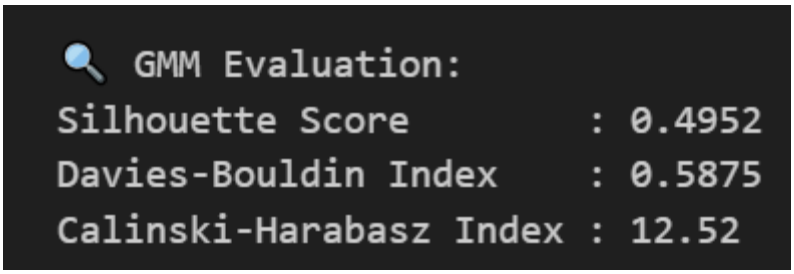


Fig.12. Regression metrics - GMM clustering

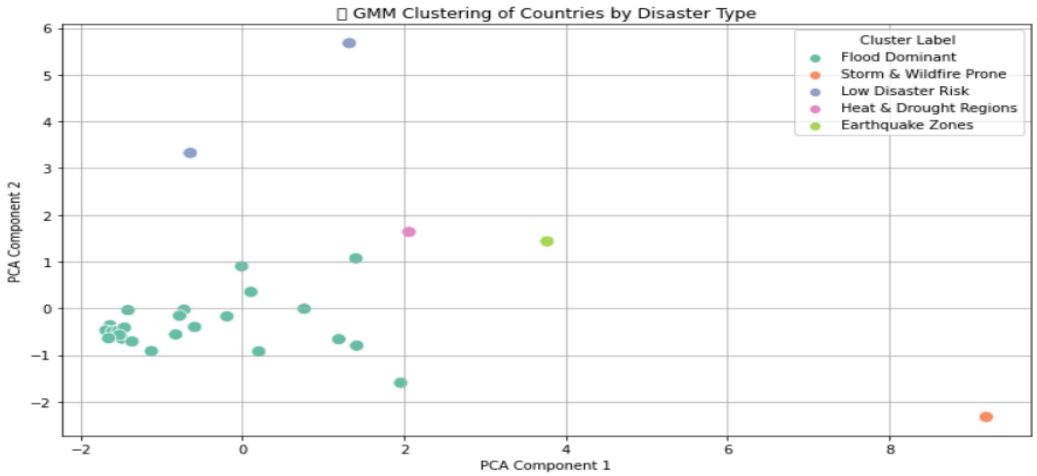


Fig.13. GMM Clustering - PCA components

Insights

- Population and area-adjusted disaster counts improved the fairness and accuracy of clustering.
- Agglomerative Clustering gave the most domain-relevant clusters with interpretable patterns.
- Clustering results offer a valuable layer for risk segmentation and policy prioritization for disaster resilience planning.

6. TOOL DEVELOPMENT

To collate the insights, model outcomes, and forecasts into a user-friendly format, a highly interactive web dashboard has been developed as a single web interface using Streamlit. This interface allows the viewer to view trends, analyze model outputs, and make forecasts on disasters and climate.

This dashboard has multiple sections, each selectable through the sidebar for navigation by users in the following ways:

Overview: It contains summary statistics like overall average global temperature, CO_2 emissions, precipitation, disaster events, and the most disaster-prone countries.

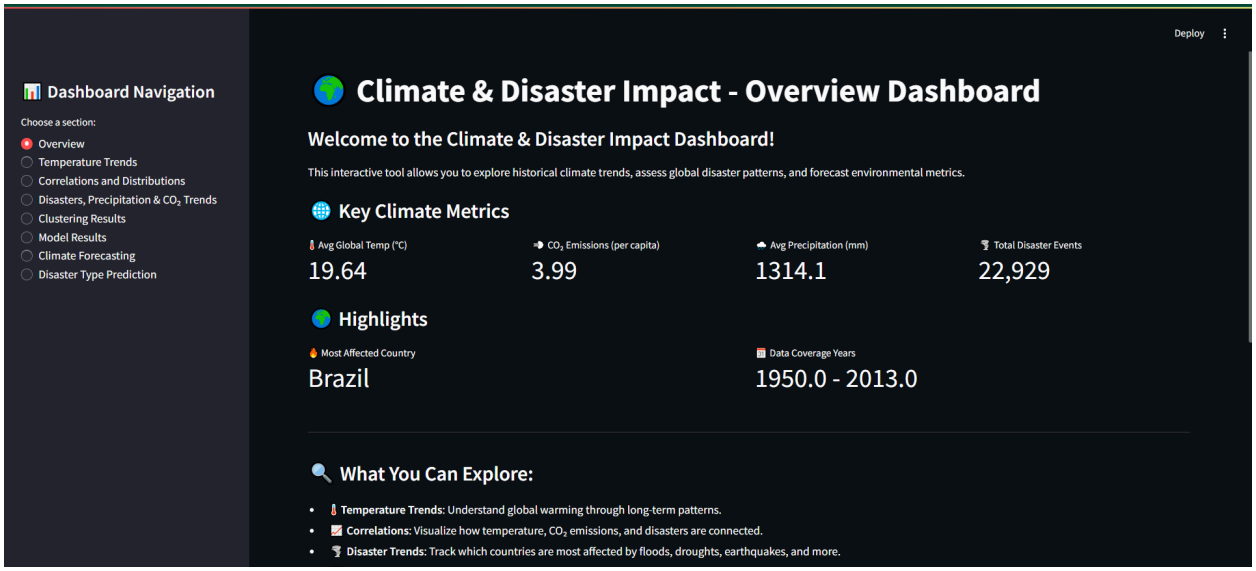


Fig.14. Dashboard - Overview

Temperature Trends: Yearly global temperature changes, uppermost hot countries, and country-specific temperature trends (U.S., for example).

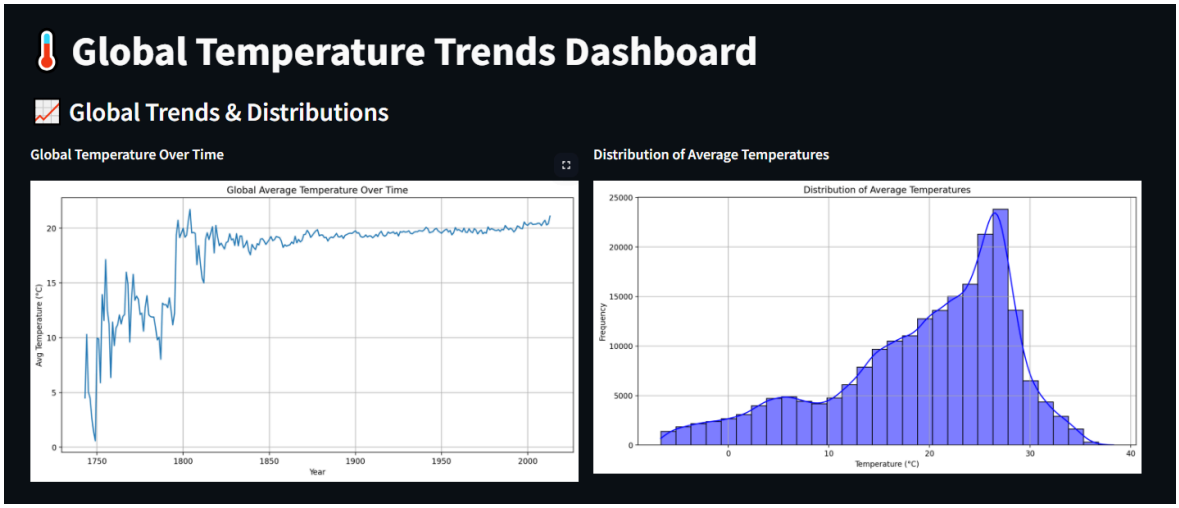


Fig.15. Dashboard - Global Temperatures

Correlations and Distributions: Heatmaps, scatter plots, and histograms are provided to investigate the interrelationship of environmental indicators and disaster data.

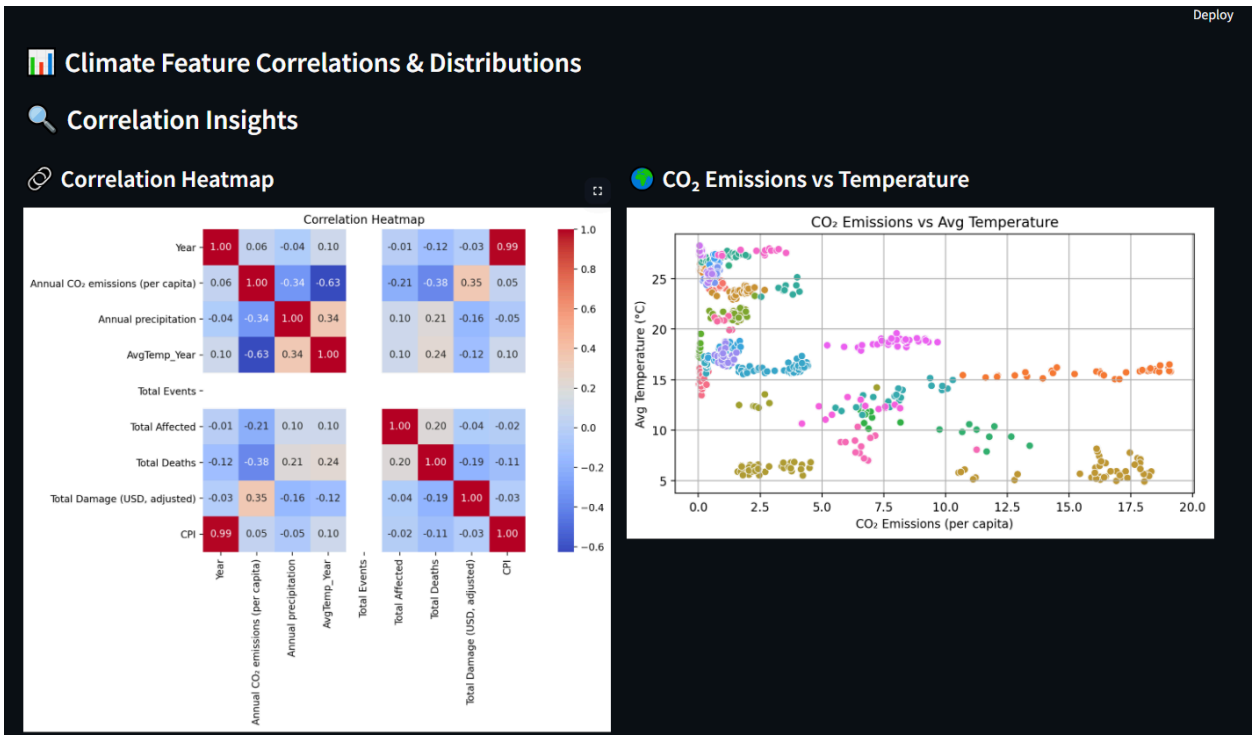


Fig.16. Dashboard - Correlation Insights

Disasters, Precipitation & CO_2 Trends: Global findings depicting how precipitation, CO_2, and phenomes evolve.

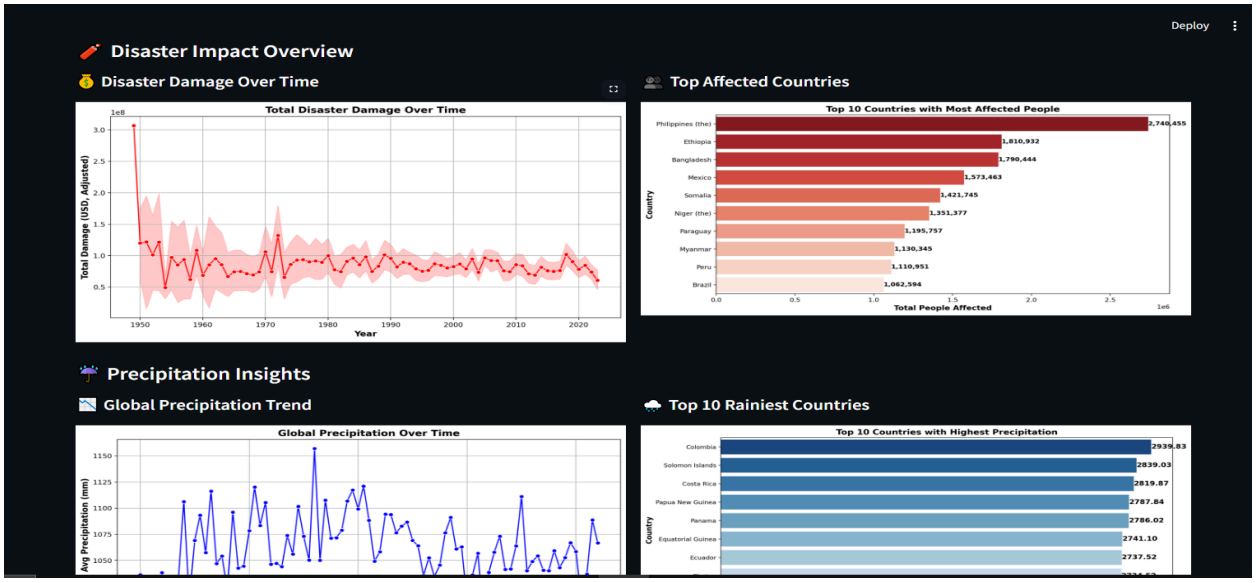
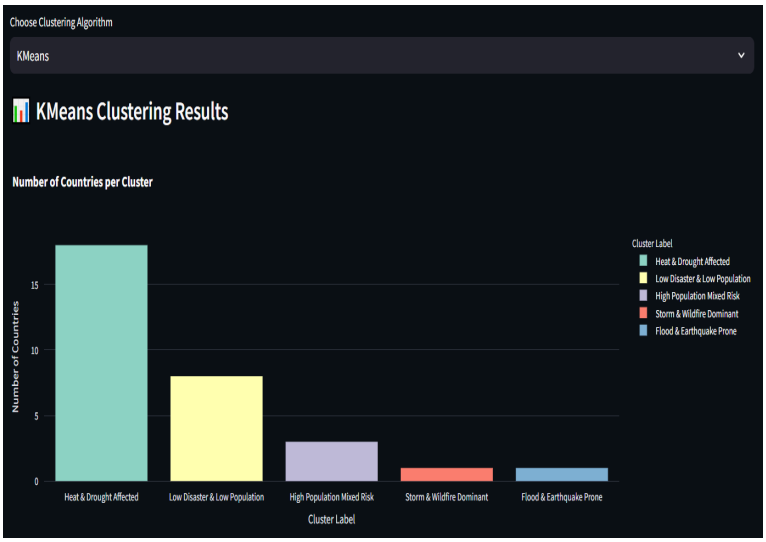


Fig.17. Dashboard - Disasters, Precipitation, and CO2 trends

Clustering Results: Country clusters are interactively visualized using K-Means, Agglomerative Clustering, and GMM, in terms of disaster impact on population.



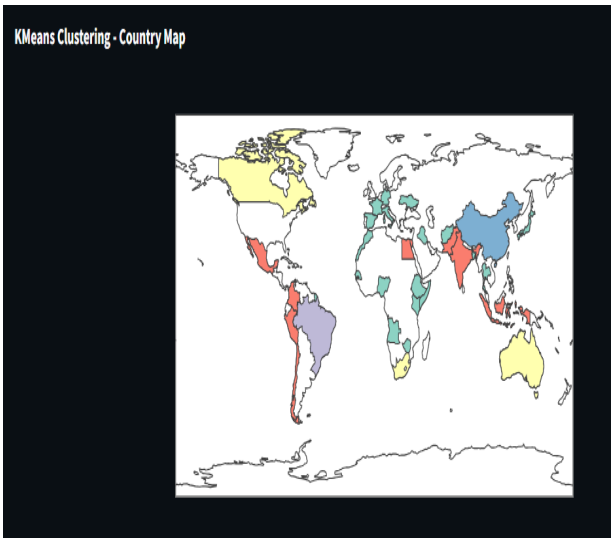


Fig.18. Dashboard - Clustering Results

Model Outcomes: Regression and classification comparisons were modeled against several algorithms for predictions, using bar charts for comparative analysis.

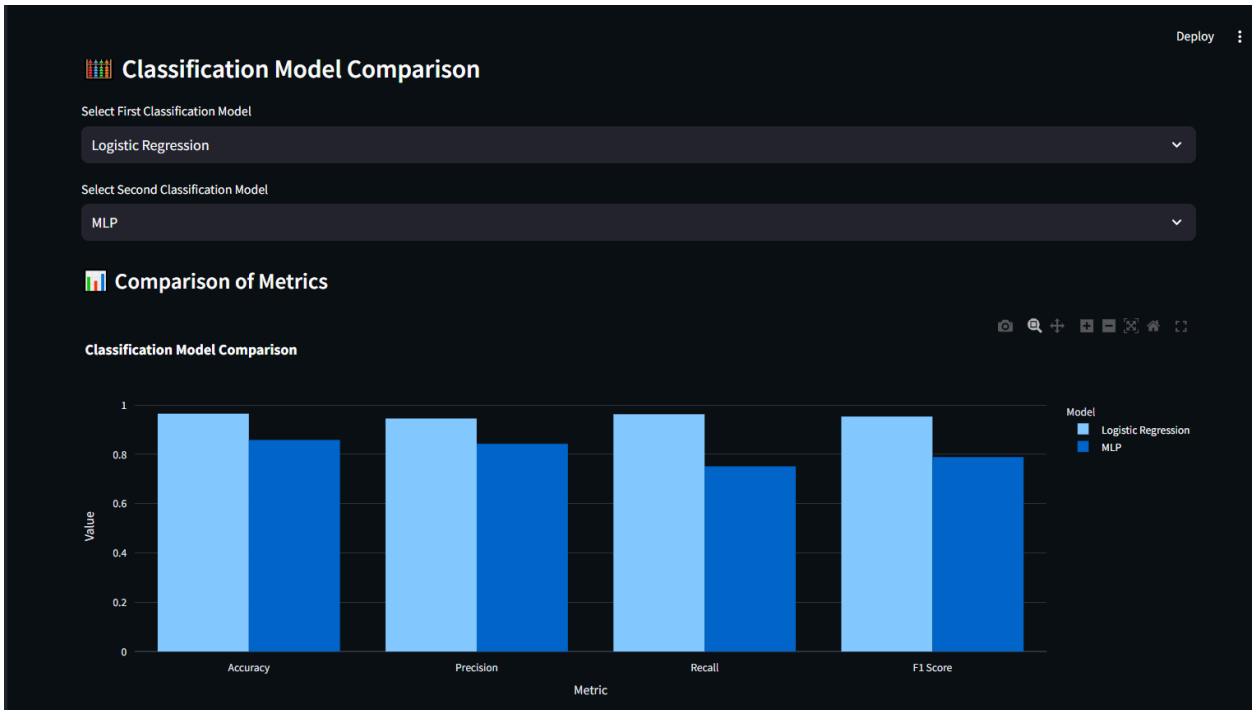


Fig.19. Dashboard - Classification and Regression results

Climate Forecasting: Users can input a country and year when temperature, precipitation, and CO_2 emissions are to be forecasted, using the regressed XGBoost model.

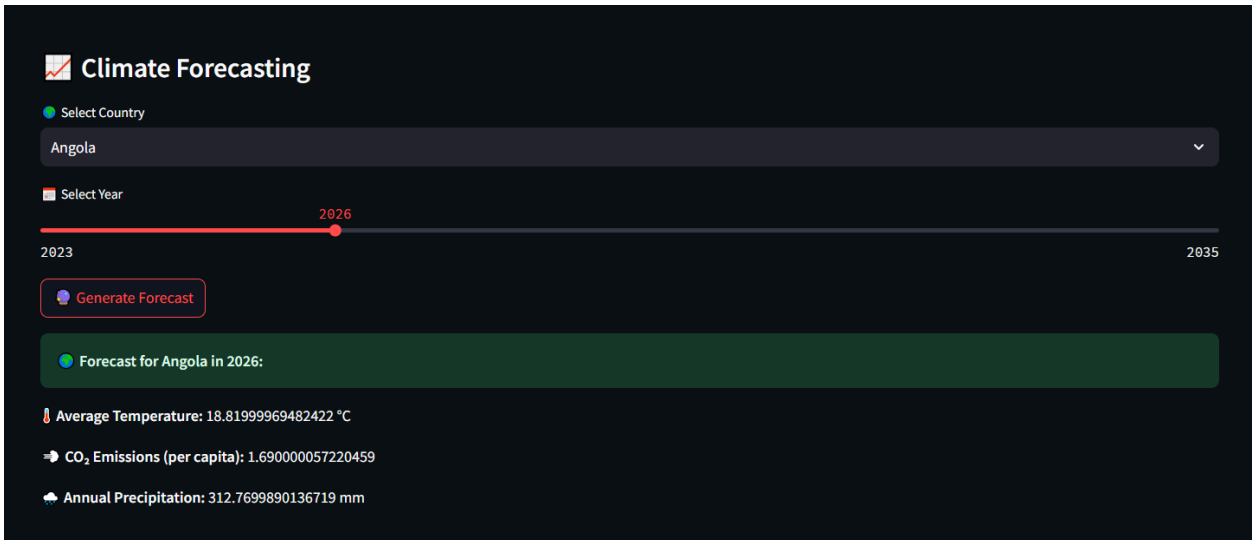


Fig.20.Dashboard - Climate Forecasting

Disaster Type Prediction: Predicts the disaster type by the model for a country and year using a Random Forest classifier based on geospatial and environmental features.

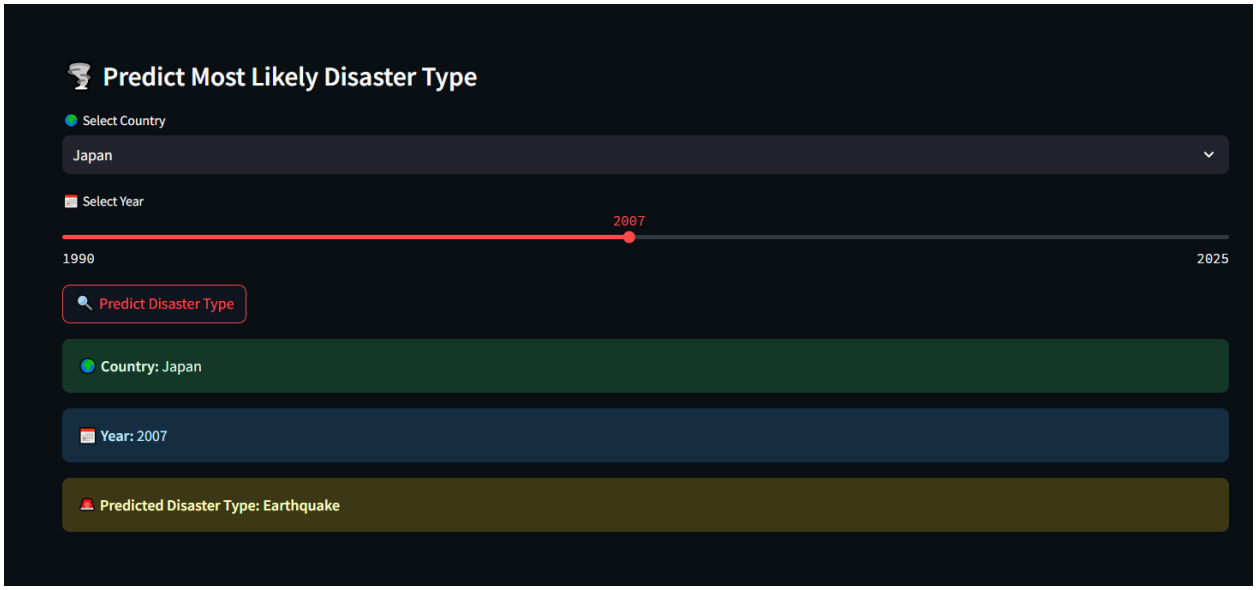


Fig.21. Disaster Type Prediction

6.2 Features Implemented

Modular visualization: To maintain the clarity of organization and reusability, all plotting and mapping events are dumped into a separate visualizations.py module.

Interactive Widgets: The dropdowns, sliders, and radio buttons were set up for user input and dynamic interaction.

Integration of Models: The trained models (Random Forest, XGBoost, MLP) were saved using joblib and loaded during runtime for real-time inference.

Dynamic Layouts: Each section for the dashboard lies within st.columns() and st.expander() blocks to retain a visually appealing, compact design.

6.3 Technical Stack

Frontend: For laying out the framework, human interaction, and for deployment purposes, Streamlit

Backend Models: The success of machine learning models was ensured using scikit-learn (for traditional ML), XGBoost (for more complexity), and LSTM using TensorFlow and Keras.

Visualizations: Seaborn; Matplotlib; Plotly; Basemap.

6.4 Insights and Impact

The tool turns complex data and ML outputs into interpretable visual insights to aid the decision-making processes of environmental policies, disaster preparedness, and climate awareness. It enables:

- Comparing the severity of disaster impact across countries
- Identifying environmental risk clusters
- Forecasting future climatic variables
- Predicting likely disasters for any region in a particular year

6.5 Limitations & Future Work

Currently, predictions are based on historical data that have undergone preprocessing. Hence, real-time or live updates are not in the system yet.

The set coordinates for countries where disaster prediction takes place would benefit from geocoding API integration for scalability (be that via Google or OpenStreetMap).

General features are being used for disaster prediction, while in the future, the integration of disaster preparedness indices, urbanization indices, or emergency response could be considered in further versions.

7. CONCLUSION

The project was able to showcase how the integration of climate data with machine learning techniques could be used to study and analyze trends for environmental changes around the globe, as well as risks of possible disasters. A comprehensive data set of climatic and socio-economic features across countries was developed through careful feature engineering and selection. The disaster impact predictions tend to perform outstandingly with the classification models we developed, especially XGBoost, whereas key climatic indicators such as temperature, CO_2 emissions, and precipitation were predicted best by Random Forest Regression.

Clustering methods, especially Agglomerative Clustering, effectively grouped countries based on disaster exposure and population-adjusted metrics, offering valuable insights for risk segmentation. Finally, I created an interactive Streamlit dashboard that makes forecasting, disaster type prediction, clustering analysis, and model comparison possible with a single tool. With this, users can find their trends, predictions, and insights through an easy-to-use interface. So, in all respects, this project outlines a scalable framework with a good capability for climate- and disaster-related analysis, and future enhancement possibilities with real-time information integration and broader geographical coverage.