INTRODUCTION TO DATA SCIENCE
MILESTONE - II
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 EDA - RECAP

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.

Feature Engineering and Feature Selection

Here, the features Year, and Average_temp_by_year are extracted from the date column and Average_temp columns and few categorical variables are changed to numerical values using label encoding. The latitude and longitude values are converted to float by removing the direction alphabet.

During the integration process, missing data was addressed by eliminating unnecessary records and filling specific columns with their respective mean values. Following this, new features were engineered, and key variables were selected to align with the requirements of the classification tasks.

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.

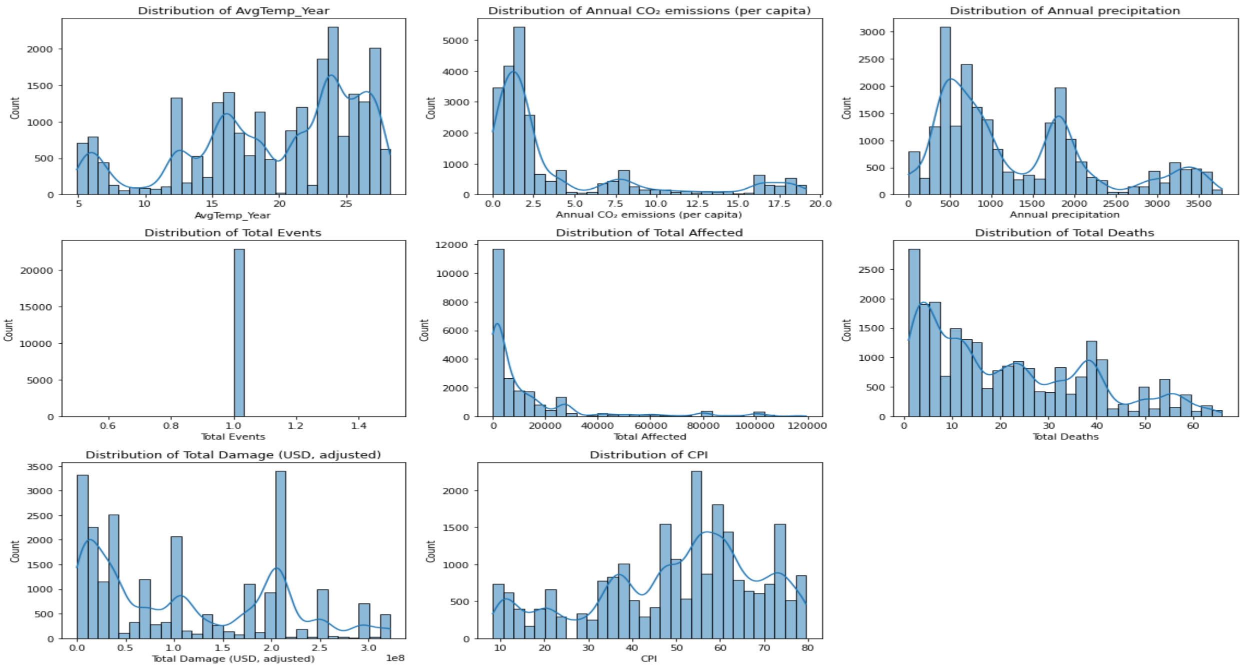


Fig.1. Histograms of different features

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.

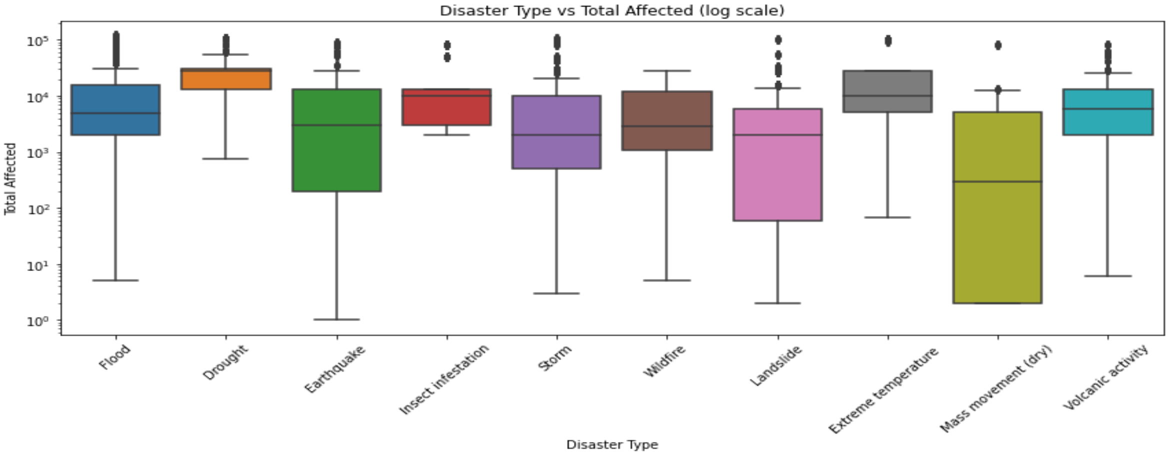


Fig.2. Boxplot regarding the disasters

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.

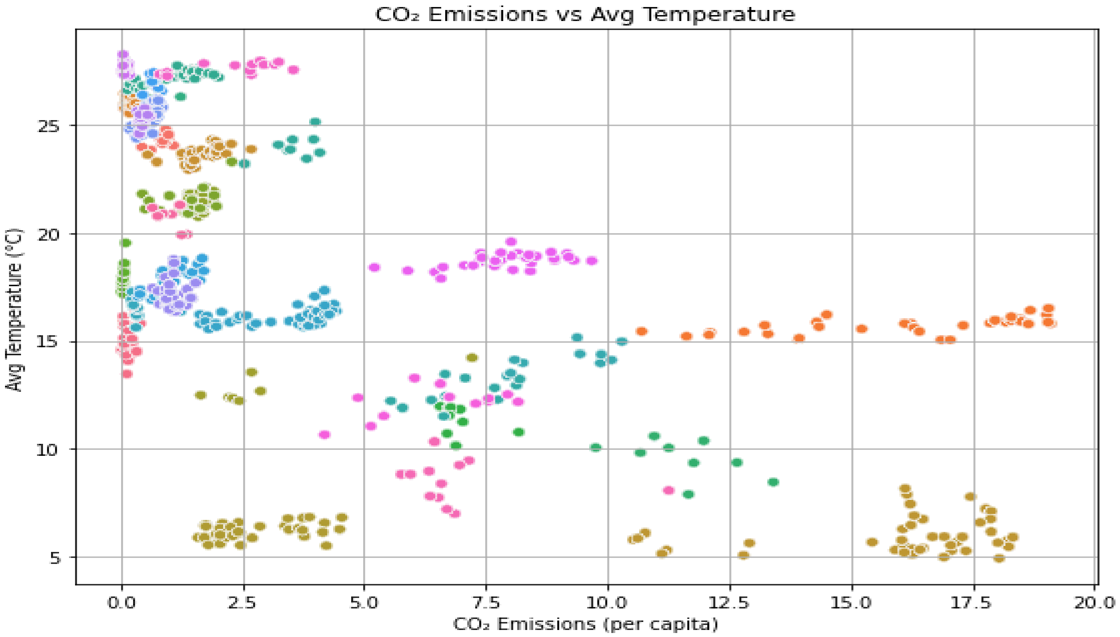


Fig.3. Average Temperature vs CO2 emissions

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. FEATURE ENGINEERING

Feature Engineering involves creating significant features from the new data which can be applied to train machine learning models for better results and accuracy.

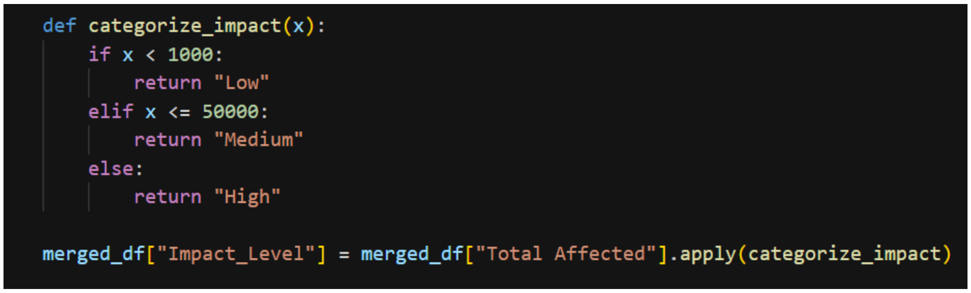
In this project, many features were engineered to enhance the environmental and climatic conditions in relation to each country to support in giving great insights and context for predictive modeling, which include the following:

4.1. Created a New Target Feature: Disaster Impact Severity

To frame the problem as a classification task, a new categorical target was engineered from the Total Affected column which includes three classes as shown below.

- Low Impact: Total affected people < 1,000
- Medium Impact: 1,000 ≤ Total affected < 50,000
- High Impact: Total affected ≥ 50,000

This new Disaster Impact Severity label allows models to learn patterns that correlate disaster impact levels with environmental, geographical, and economic features.

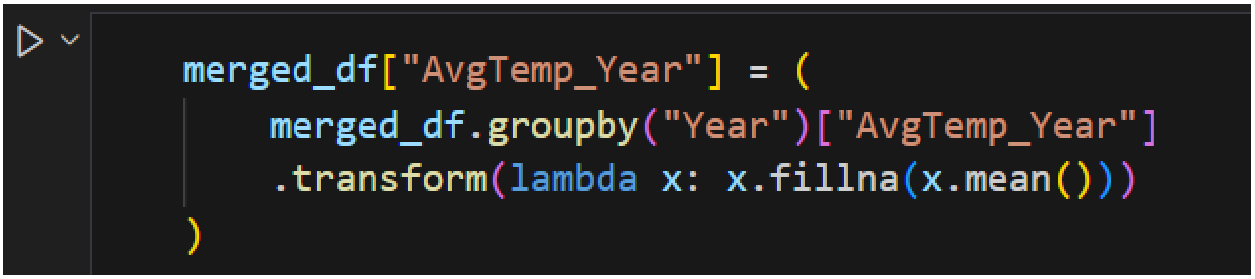


```
def categorize_impact(x):
    if x < 1000:
        return "Low"
    elif x <= 50000:
        return "Medium"
    else:
        return "High"

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

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

4.2. Aggregated Monthly Temperature into Yearly Averages (AvgTemp_Year)

The raw temperature dataset originally included monthly average temperatures for each city or region. To simplify and unify this data for country-level modeling The temperatures were grouped by country and year. The mean temperature across all months in a year was computed, resulting in a new feature: **AvgTemp_Year**. This transformation helps relate temperature trends with long-term climate effects and disasters rather than short-term variations.



```
merged_df["AvgTemp_Year"] = (
    merged_df.groupby("Year")["AvgTemp_Year"]
    .transform(lambda x: x.fillna(x.mean()))
)
```

Fig.5. Code snippet for extracting AvgTemp_Year

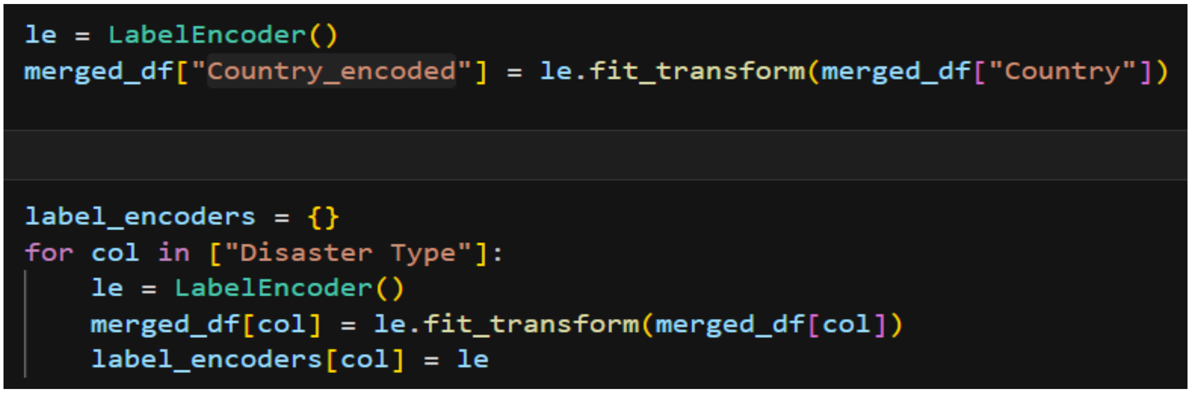
4.3. Merged Emissions and Precipitation Data with Country-Level Disaster History

Multiple datasets were combined to form a comprehensive view of each country's environmental and disaster context. In that process, Emissions and Precipitation Data are merged with Country-Level Disaster History on the basis of year. This forms a comprehensive view of each country's environmental and disaster context

This enriched the disaster dataset with environmental factors that may have a certain effect on or be affected by disasters (e.g., High-rainfall floods, emissions contributing to climate events). Disaster records (number of deaths, damage cost, etc.) were matched with these environmental metrics from the same year and country.

4.4. Encoding categorical columns

As Machine Learning models work well with numerical data. Several string based columns like country and disaster type are encoded into numerical columns using label encoder. This enables clustering, classification, and regression models to interpret these fields without bias toward any alphabetical or arbitrary ordering.

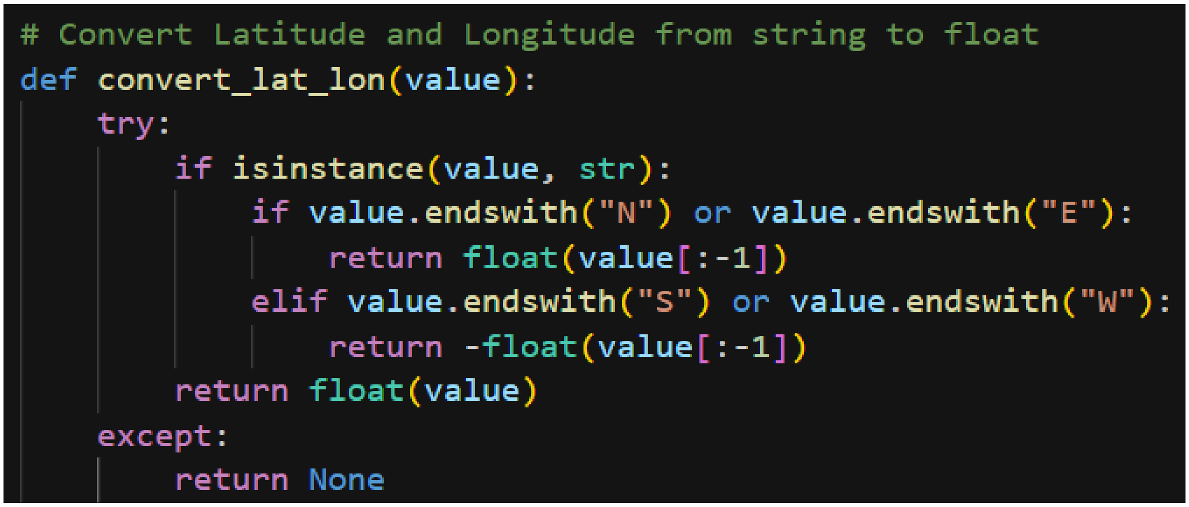


```
le = LabelEncoder()
merged_df["Country_encoded"] = le.fit_transform(merged_df["Country"])
```

```
label_encoders = {}
for col in ["Disaster Type"]:
    le = LabelEncoder()
    merged_df[col] = le.fit_transform(merged_df[col])
    label_encoders[col] = le
```

Fig.6. Code snippet for using label_encoder

The conversion of geospatial coordinates (Latitude, Longitude) from string formats (for example, "34.56N") into numerical values was done using a custom convert_lat_lon() function. Missing values were handled by dropping invalid entries.



```python
# Convert Latitude and Longitude from string to float
def convert_lat_lon(value):
    try:
        if isinstance(value, str):
            if value.endswith("N") or value.endswith("E"):
                return float(value[:-1])
            elif value.endswith("S") or value.endswith("W"):
                return -float(value[:-1])
        return float(value)
    except:
        return None
```

Fig.7. Code snippet for converting latitude and longitude

5. FEATURE SELECTION

Feature Selection is the process of identifying the most important subset of features from a larger dataset aiming for better performance reducing the computational cost.

5.1. Correlation Analysis for Feature Selection

- Before building predictive models, a correlation heatmap was used to examine the linear relationships between numerical features in the merged dataset. Identifying 'Highly correlated features that would induce redundancy or multicollinearity'. Features exhibiting low or zero correlation with the target variable implying possible less predictive powers features.
- Insights:
- Usually, total affected, total damage, and total deaths are correlated, as they are all measures of disaster intensity.
- Environmental factors such as AvgTemp_Yr, Annual precipitation, and CO2 emissions were moderately correlated among themselves but still important parameters for predicting disaster-induced outcomes.
- Categorical encodings Country and Disaster Type showed very little linear correlation with numerical features, which is expected, yet should not decrease their significance in tree-based models.
- In this context, redundant features were taken into account to be removed or transformed, the meaningful ones were retained to be included in modeling.

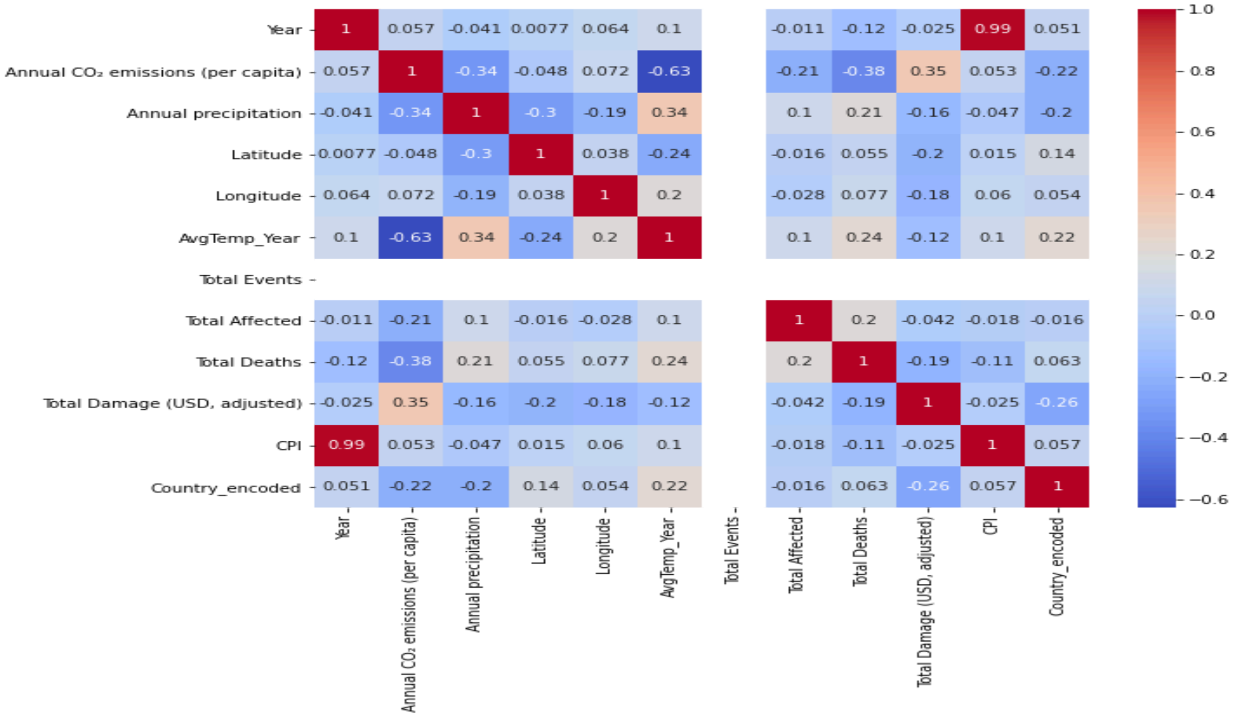


Fig.8. Correlation between features

5.2. Feature Importance (Random Forest)

The built-in feature importance from Random Forest itself was used to find out the more important parts that influenced mostly the disaster impact classification model. The feature importance plot ranks all input variables in terms of how much each one contributed to the decision-making processes of the model. They are then used to calculate the distinction of these features based on how often and how well these features are used for splitting among all the trees in the forest.

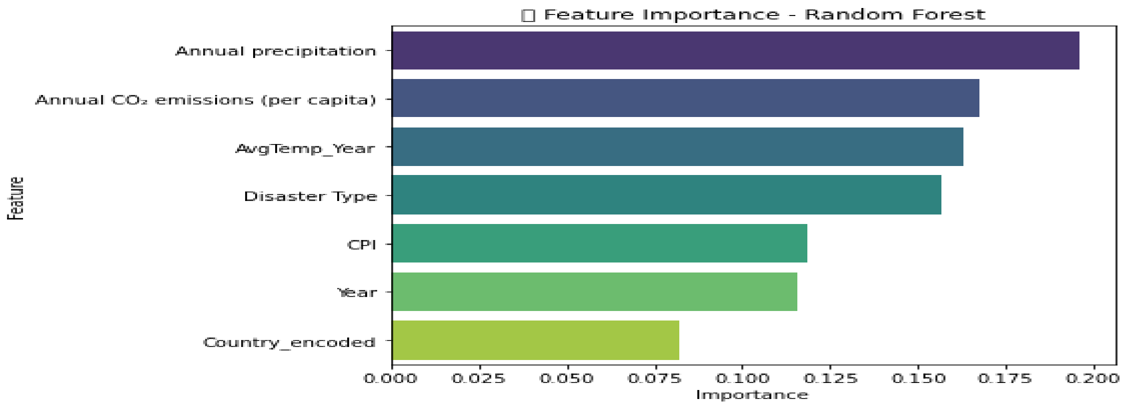


Fig.9. Feature Importance - Random Forest

Insights

- Climate-related features such as AvgTemp_Year, Annual CO_2 emissions, and Annual precipitation proved to be the most influential and demonstrated how environmental conditions affect the impact of disasters.
- Total Affected and Total Damage (USD, adjusted) emerged also as influential features, which aligns well with the classification objective of determining disaster severity.
- Socio-economic indicators like CPI (Consumer Price Index) and Country help explain the resilience or vulnerability of a region.

5.3. 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. The idea was to obtain a small number of transformed components that would account for most of the variance present in the original dataset. A cumulative-explained variance plot was then produced in this analysis describing that:

- 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.

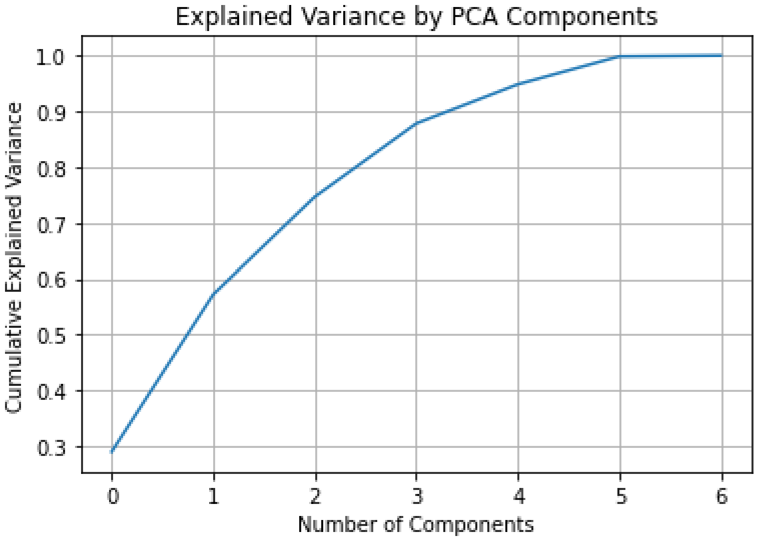


Fig.10. PCA explaining 90% variance

6. DATA MODELING

6.1 Data Splitting

Reliably building machine learning models; thus, data is split deliberately into two, where 80% of it is designated for training while 20% is reserved for testing or validation. This action is called data splitting, and it allows the model to learn from almost all of the data but test it on what it has never seen before. The training subset is the one being used to fit the model and probably even optimize its parameters while testing is meant to develop the model's ability to generalize into inputs of new, unseen examples. This somehow can detect random issues such as overfitting or underfitting and give an expected reality about the whole predictive performance of the model.

6.2. Model Training

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

- A classification task for predicting the severity of disaster impact 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.

Classification Task

Objective: predicting the severity of disaster impact for a given country-year-disaster combination.

Target Variable:

A new categorical feature named Disaster Impact Severity was created by binning the Total Affected column into three classes

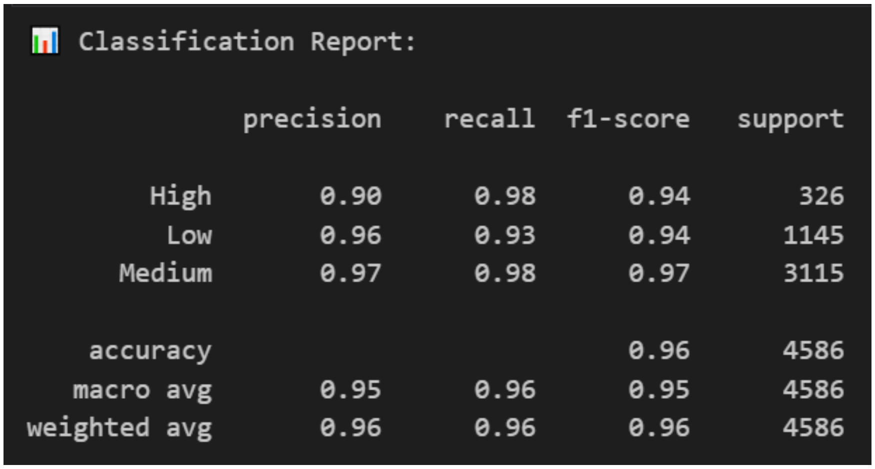
- Low Impact: Total Affected < 1,000
- Medium Impact: 1,000 ≤ Total Affected < 50,000
- High Impact: Total Affected ≥ 50,000

Models Implemented

1. Random Forest
2. MLP - Neural Network
3. XGBoost

Random Forest

Random Forest is a powerful and widely used ensemble learning algorithm primarily used for classification and regression tasks. It operates by constructing multiple decision trees during training time and outputs the mode (classification) or mean (regression) of the predictions from individual trees. This model is trained on the splitted data and evaluation results are as follows



```
📊 Classification Report:

              precision    recall  f1-score   support

        High       0.90      0.98      0.94       326
         Low       0.96      0.93      0.94      1145
      Medium       0.97      0.98      0.97      3115

    accuracy                           0.96      4586
   macro avg       0.95      0.96      0.95      4586
weighted avg       0.96      0.96      0.96      4586
```

Fig.11. Classification Report - Random Forest

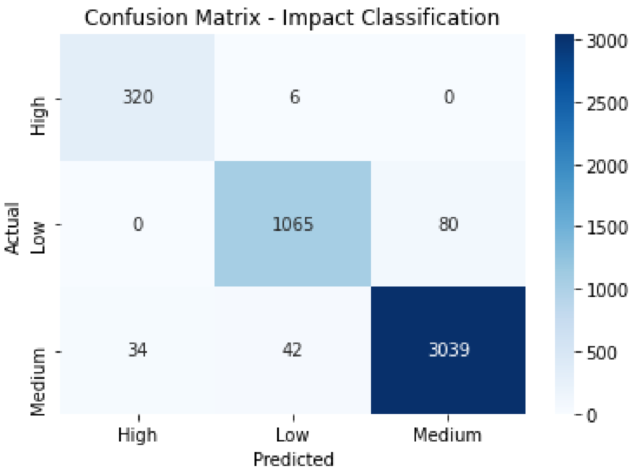
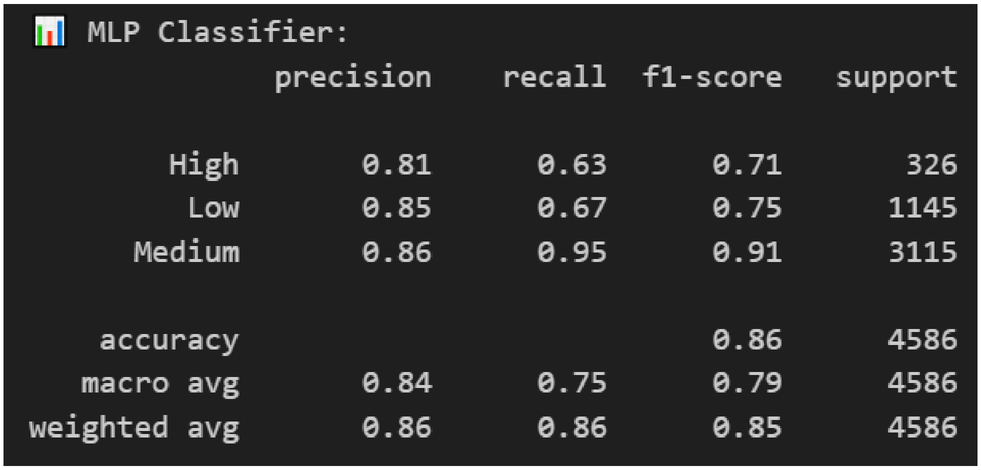


Fig.12. Confusion Matrix - Random Forest

The model performs very well for predicting Medium impact events with a high true positive rate (3039 correct). There is some confusion between Low and Medium classes (80 and 42 misclassified). High impact disasters are well predicted too, with only 6 being misclassified.

MLP - Neural Network

A Multilayer Perceptron (MLP) is a type of artificial neural network that consists of an input layer, one or more hidden layers, and an output layer. It is a supervised learning algorithm capable of learning non-linear mappings between input features and target variables by passing data through a series of layers where each neuron applies a learned weight and activation function.



```
📊 MLP Classifier:
              precision    recall  f1-score   support

        High       0.81      0.63      0.71       326
         Low       0.85      0.67      0.75      1145
      Medium       0.86      0.95      0.91      3115

    accuracy                           0.86      4586
   macro avg       0.84      0.75      0.79      4586
weighted avg       0.86      0.86      0.85      4586
```

Fig.13. Classification Report - MLP

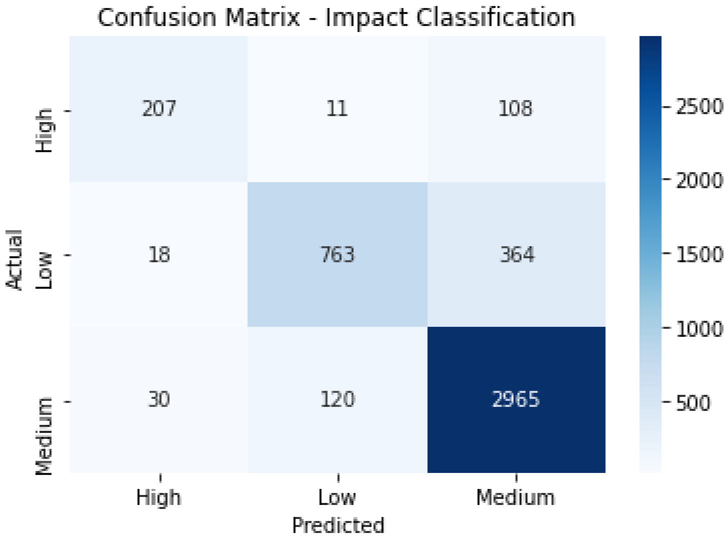
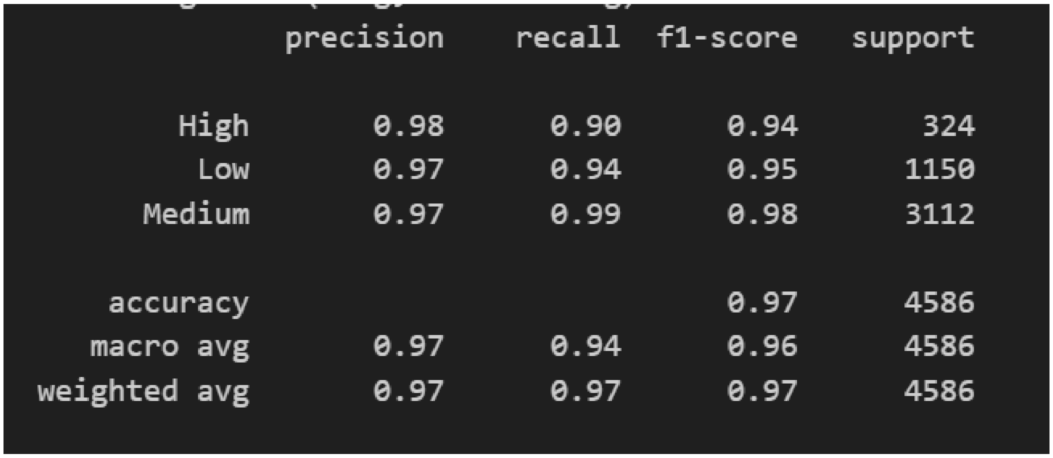


Fig.14. Confusion Matrix - MLP

The model performs best for medium impact disasters, as seen from the highest correct predictions (2965). Low impact and high impact categories have more confusion, especially misclassification into the medium class.

XGBoost

XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable machine learning algorithm based on gradient boosting. It builds an ensemble of decision trees in a sequential manner, where each new tree attempts to correct the errors made by the previous ones. XGBoost is known for its regularization capabilities (L1 and L2), which help in reducing overfitting and improving generalization.



```
              precision    recall  f1-score   support

        High       0.98      0.90      0.94       324
         Low       0.97      0.94      0.95      1150
      Medium       0.97      0.99      0.98      3112

    accuracy                           0.97      4586
   macro avg       0.97      0.94      0.96      4586
weighted avg       0.97      0.97      0.97      4586
```

Fig.15. Classification Report - XGBoost

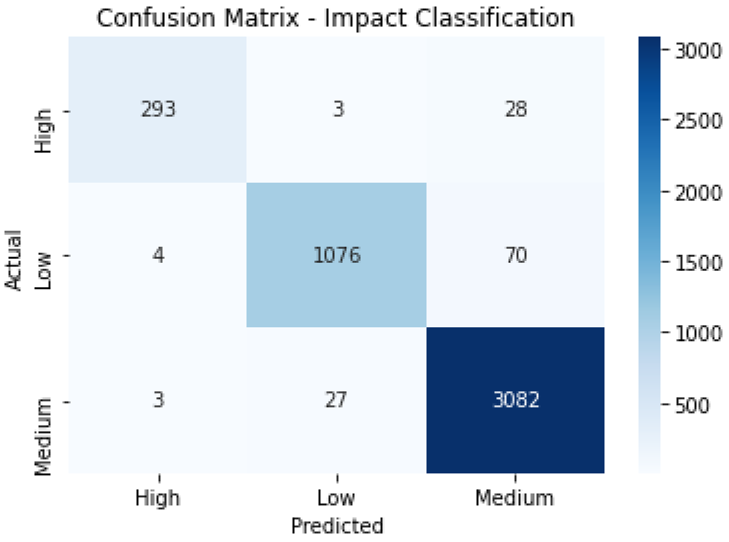


Fig.16. Confusion Matrix - XGBoost

Overall Accuracy is high, especially for the Medium and Low classes. The classifier performs exceptionally well on Medium impact disasters, with very few misclassifications. There are minimal misclassifications between High and Low, which indicates good model separation between extreme cases.

Best Model

XGBoost is the best model with best precision and recall overall. It is having fewer false positives and negatives and also with better accuracy among the three models which is around 97%.

Regression Task

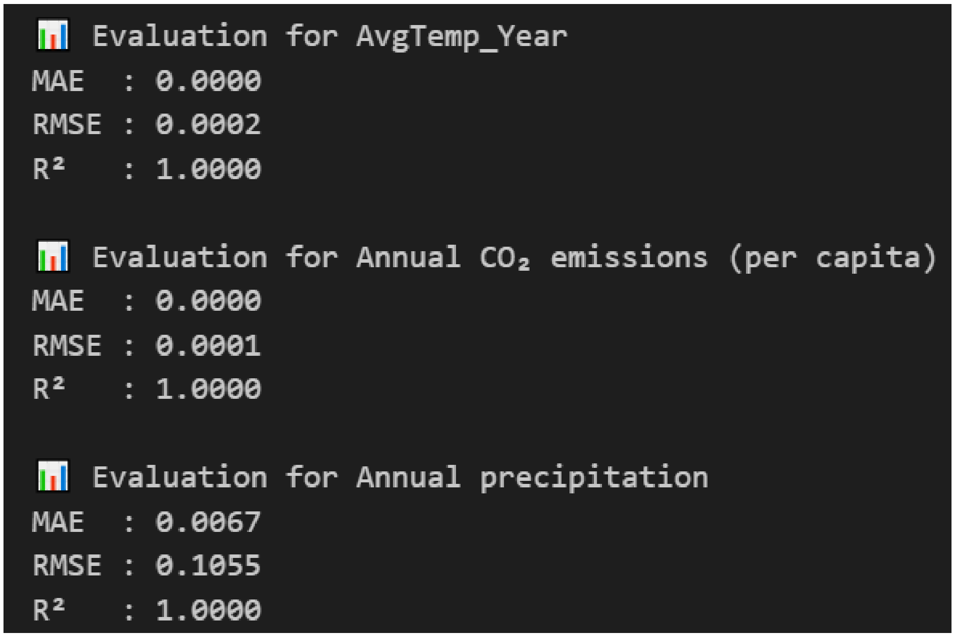
Objective: Forecasting future climate indicators like temperature, co2 emissions and precipitation based on country and year.

Models Implemented

- Random Forest Regression
- LSTM
- XGBoost

Random Forest Regression

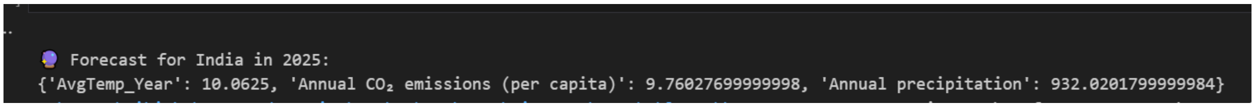
Random Forest Regression is an ensemble learning method used for predicting continuous numerical values. It operates by constructing multiple decision trees during training and averaging their outputs to produce more accurate and stable predictions. This technique helps reduce overfitting and improves generalization.



```
📊 Evaluation for AvgTemp_Year
MAE  : 0.0000
RMSE : 0.0002
R²   : 1.0000

📊 Evaluation for Annual CO₂ emissions (per capita)
MAE  : 0.0000
RMSE : 0.0001
R²   : 1.0000

📊 Evaluation for Annual precipitation
MAE  : 0.0067
RMSE : 0.1055
R²   : 1.0000
```

Fig.17. Evaluation Metrics - Random Forest Regression

Forecasting Results

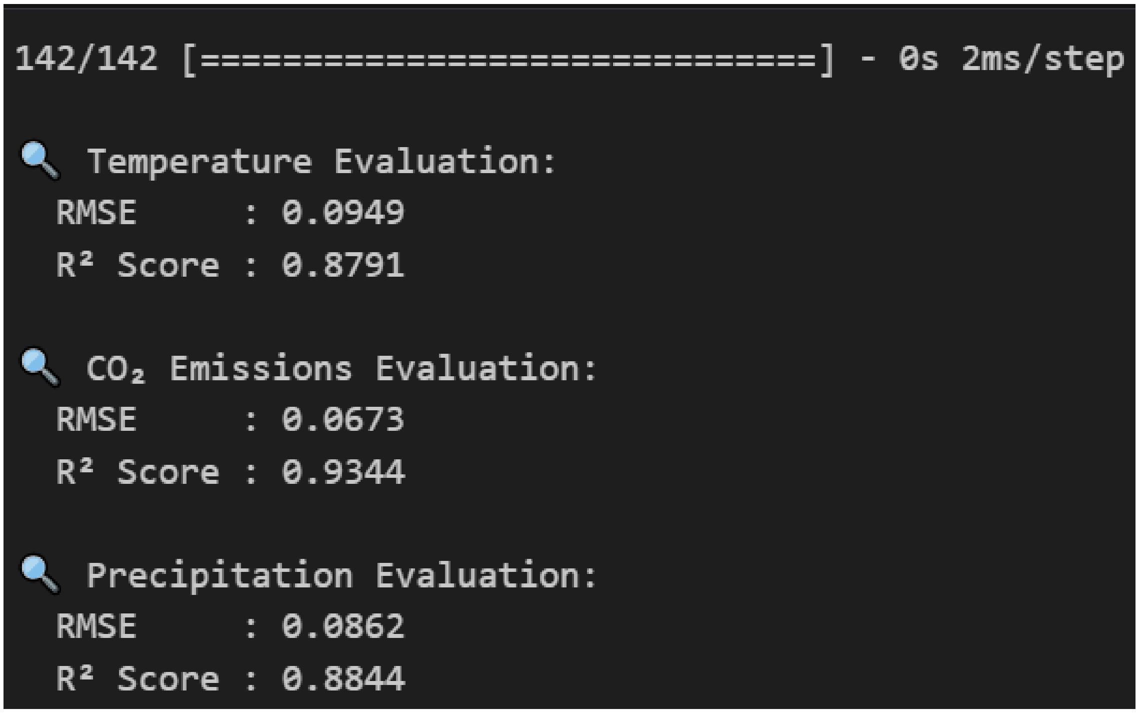


```
🔮 Forecast for India in 2025:
{'AvgTemp_Year': 10.0625, 'Annual CO₂ emissions (per capita)': 9.76027699999998, 'Annual precipitation': 932.0201799999984}
```

Fig.18. Forecast Results - Random Forest Regression

LSTM

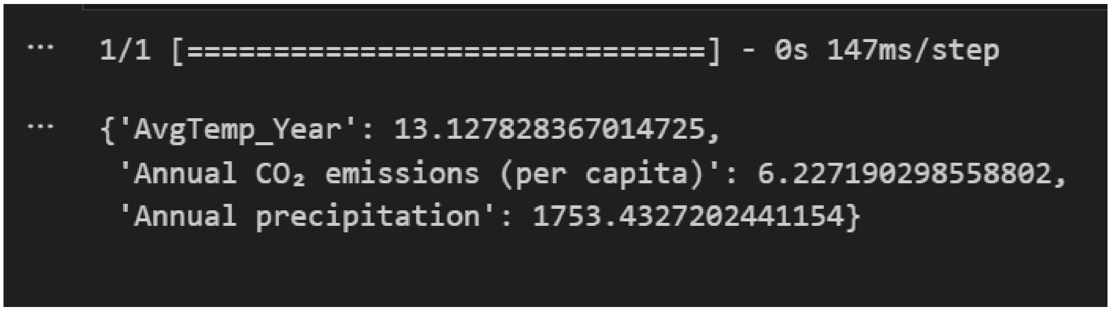
LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) designed to effectively capture long-term dependencies in sequential data. Unlike traditional RNNs, which struggle with learning from long sequences due to the vanishing gradient problem, LSTM networks use memory cells and gating mechanisms (input, forget, and output gates) to retain and control the flow of information over time.



```
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.19. Evaluation Metrics - LSTM

Forecasting Results

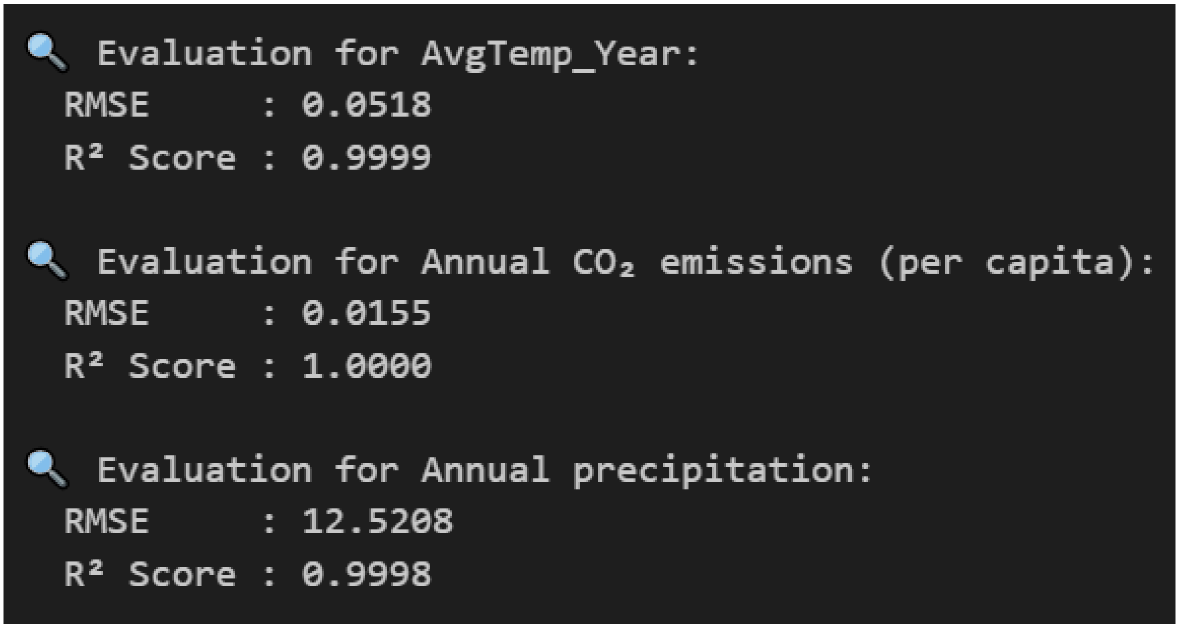


```
···  1/1 [==============================] - 0s 147ms/step

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

Fig.20. Forecast Results - LSTM

XGBoost

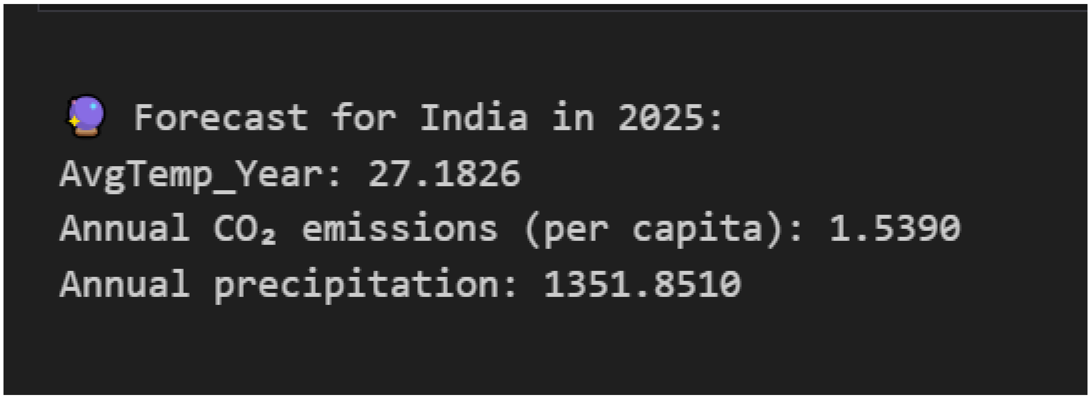
XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable machine learning algorithm based on gradient boosting. It builds an ensemble of decision trees in a sequential manner, where each new tree attempts to correct the errors made by the previous ones. XGBoost is known for its regularization capabilities (L1 and L2), which help in reducing overfitting and improving generalization.



```
🔍 Evaluation for AvgTemp_Year:
  RMSE     : 0.0518
  R² Score : 0.9999

🔍 Evaluation for Annual CO₂ emissions (per capita):
  RMSE     : 0.0155
  R² Score : 1.0000

🔍 Evaluation for Annual precipitation:
  RMSE     : 12.5208
  R² Score : 0.9998
```

Fig.21. Evaluation Metrics - LSTM

Forecasting Results



```
🔮 Forecast for India in 2025:
AvgTemp_Year: 27.1826
Annual CO₂ emissions (per capita): 1.5390
Annual precipitation: 1351.8510
```

Fig.22. Forecast Results - XGBoost

Best Model

Based on all the models, Random Forest is the best model. This model has near-zero MAE and RMSE for all targets. It also has perfect R^2 (1.0) indicating accurate predictions and consistent performance across all three targets that includes temperature, co2 emissions and precipitation.

7. CONCLUSION

This project successfully integrates climate and disaster-related datasets to forecast future environmental trends and classify disaster impacts. Through meticulous data preprocessing, feature engineering, exploratory data analysis, and advanced machine learning modeling, the study provides insightful patterns about climate change and disaster vulnerability across countries.

Here, classification models predict damage levels if any disasters happen while regression analyzes the data from where you can forecast the climate predictors like temperature, co2 emissions, and precipitation of its specific country.

Overall, this project brings data science techniques together with real-world datasets to contribute to disaster preparedness, climate awareness, and policymaking . This is an infinitely expandable pipeline for analysis of complex environmental data with further actionable insights.