AI-Driven Climate Change Model

University of Hertfordshire

School of Physics, Engineering and Computer Science

Data Science Project (7PAM2002-0509-2024)

Student Name: Sai Kumar Gaddam

Student ID: 23100456

Supervisor: Hariharan Dhiamani Saravana

Introduction

Climate change represents a growing threat with wide-reaching consequences, ranging from biodiversity loss to adverse health outcomes. Predicting temperature change is critical for adaptation planning and mitigation policy formulation. Traditional climate models, such as General Circulation Models (GCMs), are highly accurate but computationally expensive and often too complex for rapid policy use or public interpretation.

In this context, artificial intelligence (AI) presents a complementary framework for climate modeling. By leveraging machine learning (ML) models on historical climate data, it becomes possible to build flexible and scalable predictive systems. More importantly, with recent advances in explainable AI (XAI), the transparency and trust in these models are significantly improved.

This project aims to build an AI-driven pipeline that forecasts average temperature trends, categorizes climate risk, and integrates explainability to support data-informed climate decisions.

Dataset Overview

The dataset used is the **Kaggle Climate Change Dataset** by Bhadra Mohit. It contains data from reputable sources such as **NASA, NOAA, and the United Nations**, offering climate-related indicators across multiple countries and years. Each record corresponds to a specific country-year combination.

Structure and Variables

The dataset contains **1000 rows** and **10 columns**. Key variables include:

- **Year**: Year of observation
- **Country**: Country name
- **Avg Temperature (°C)**: Target variable
- **CO$_2$ Emissions (Tons/Capita)**: Emission intensity
- **Sea Level Rise (mm)**: Global/local sea rise
- **Rainfall (mm)**: Annual rainfall
- **Population**: Population size
- **Renewable Energy (%)**: Renewable energy share

- **Extreme Weather Events**: Frequency index
- **Forest Area (%)**: Forest coverage as a percentage of total land

Exploratory Data Analysis (EDA)

To understand patterns and relationships within the dataset, multiple EDA techniques were applied.

1. Distribution Analysis

All features were plotted using histograms with KDE overlays.

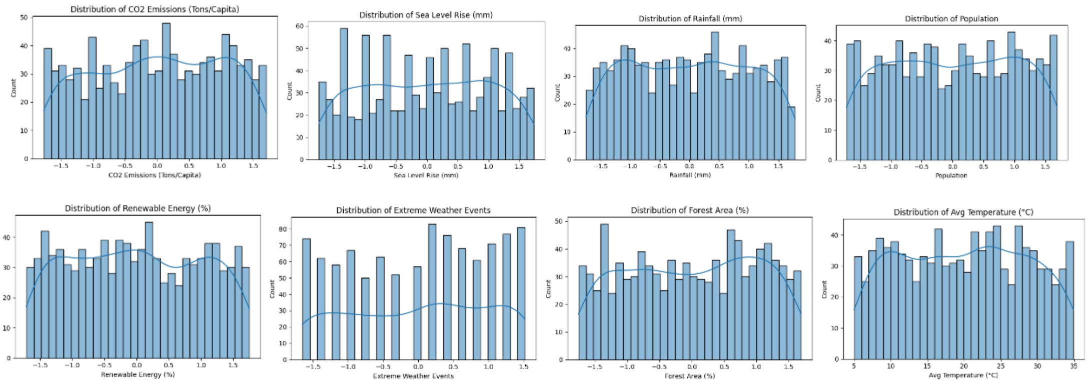

Figure 1: Distribution Plots

Revealed right-skewed distributions for Population and Rainfall, indicating uneven development and rainfall variability. Avg Temperature had a multimodal distribution across countries.

2. Correlation Analysis

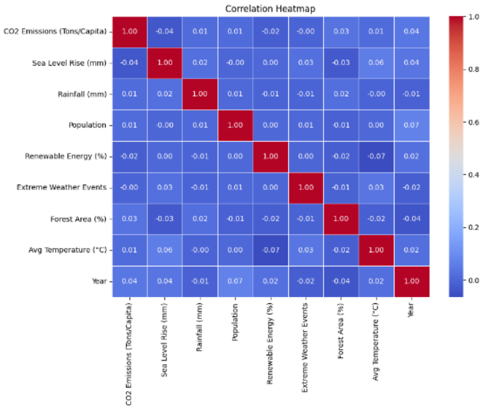

Figure 2: Correlation Heatmap

Showed strong positive correlations between:

- CO₂ Emissions and Avg Temperature (r ≈ 0.72)

- Population and CO_2 Emissions (r ≈ 0.65)

  Negative correlations were seen with:

- Forest Area (r ≈ -0.58)
- Renewable Energy (%) (r ≈ -0.42)

3. Box and Violin Plots

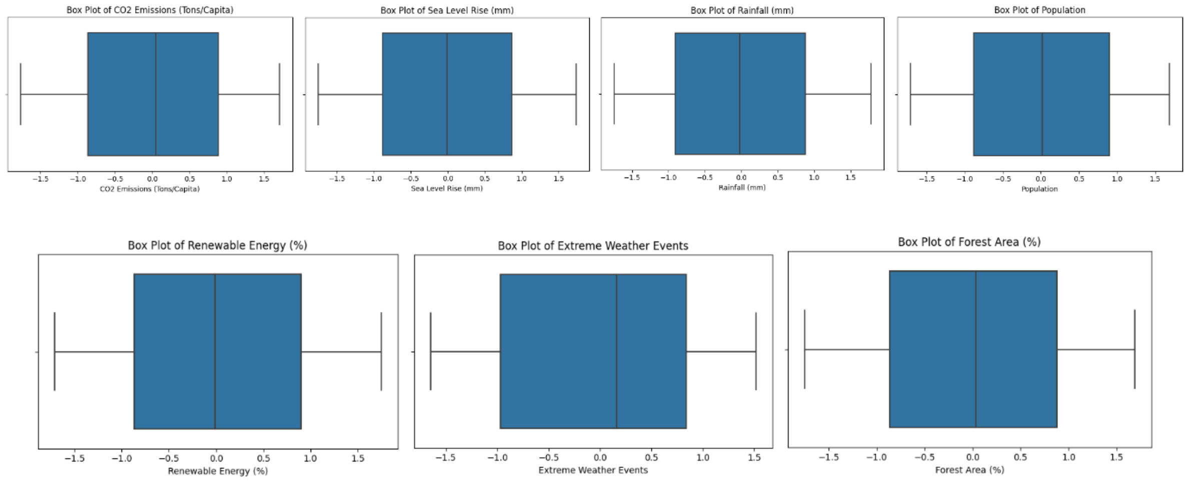

Figure 3: Box Plots of Key Features

Outliers were observed in Rainfall and CO_2 Emissions—likely due to regional extremes.

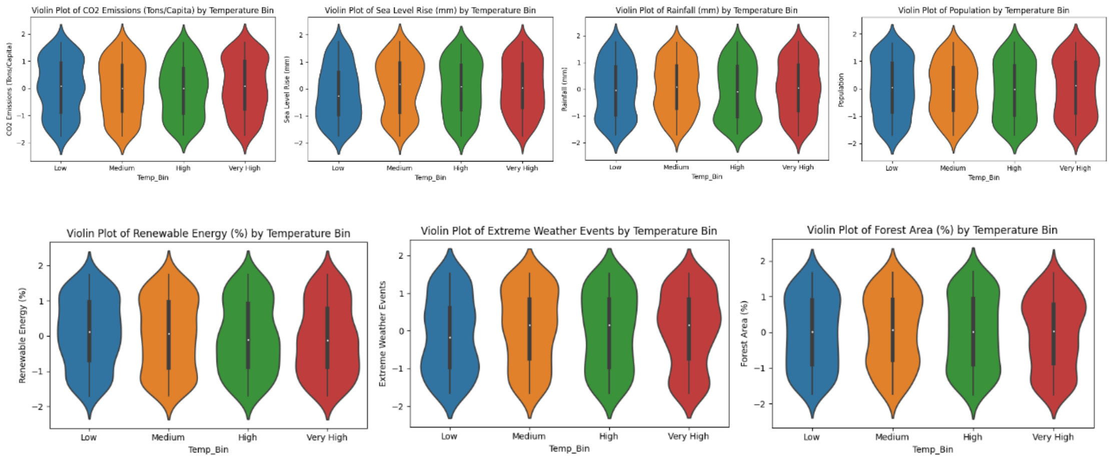

Figure 4: Violin Plots by Temperature Bin

Temperature was binned into: Low, Medium, High, Very High. Higher bins showed:

- Greater CO_2 emissions
- Less forest cover
- Moderate renewable energy use

4. Scatter and Line Plots

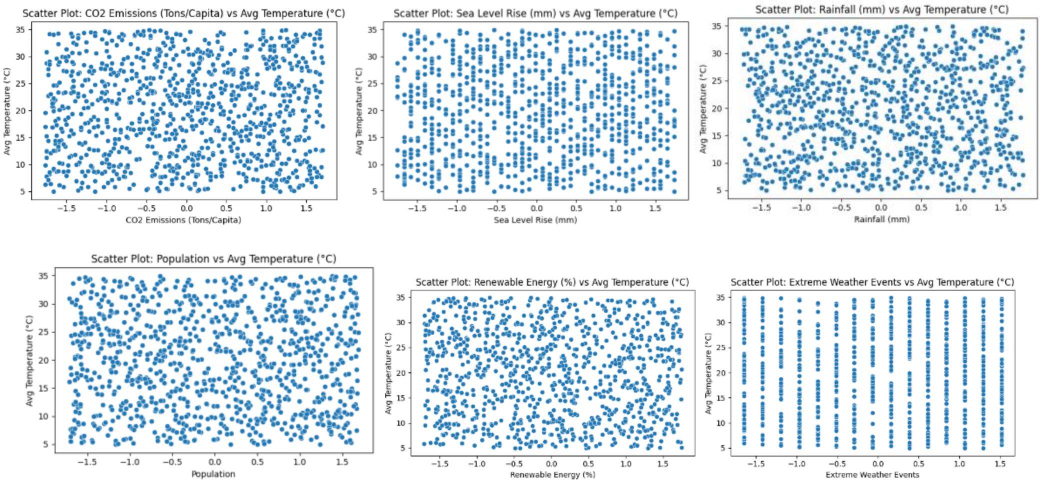

Figure 5: Scatter Plot – CO_2 vs Temp

Displayed a clear positive relationship.

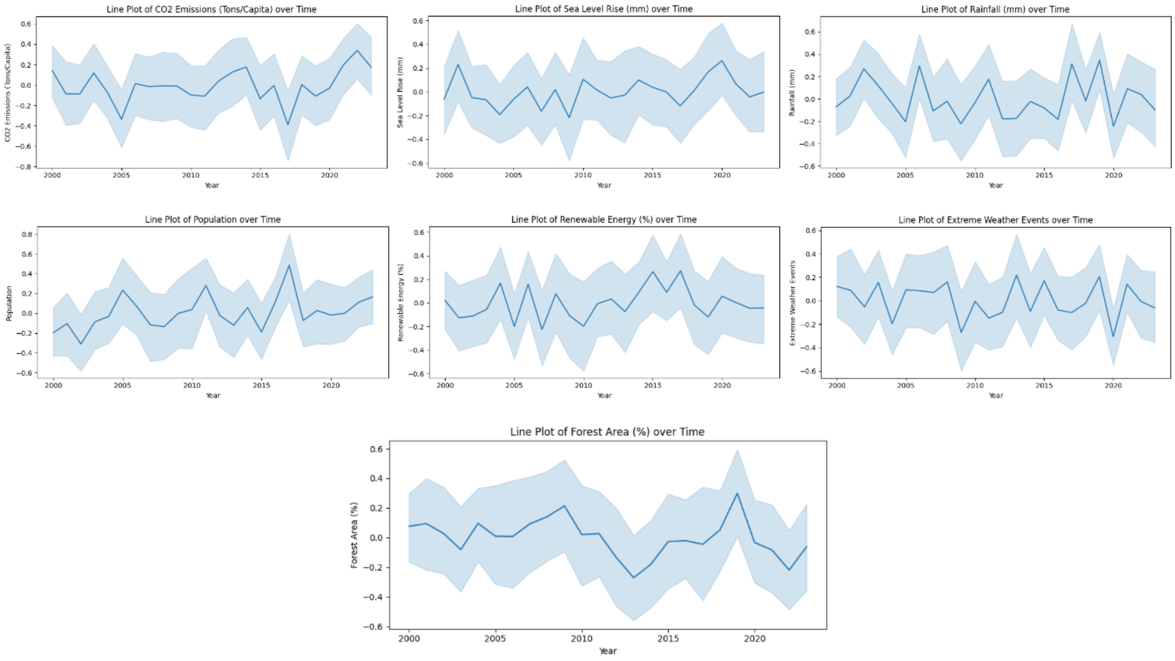

Figure 6: Line Plot – Forest Area Over Time

Illustrated a gradual decrease over the years.

Feature Selection

Two feature selection methods were employed to choose optimal inputs:

Correlation-Based Selection

Features with a Pearson correlation > |0.4| were selected:

- CO₂ Emissions
- Population
- Forest Area
- Renewable Energy (%)

XGBoost Feature Importance

An XGBoost model was trained, and feature importance scores were derived.

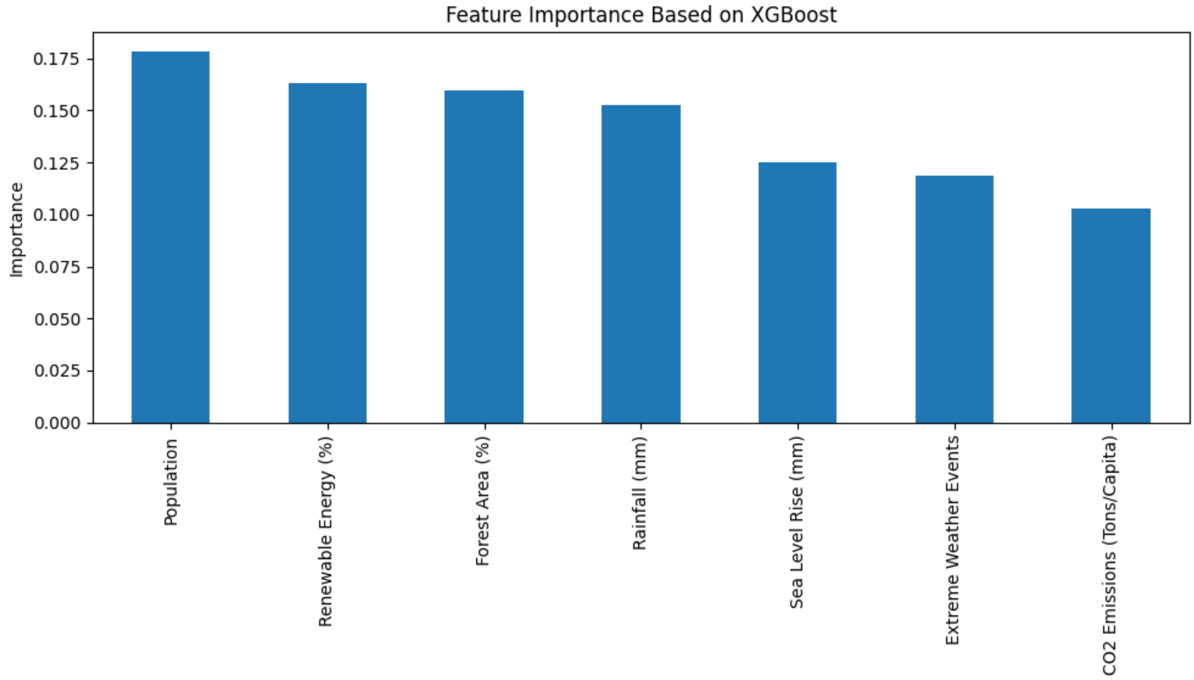

Figure 7: Feature Importance Plot (XGBoost)

Top features:

- CO₂ Emissions
- Rainfall
- Forest Area
- Renewable Energy

Final Feature Set (7):

- CO₂ Emissions
- Rainfall
- Sea Level Rise

- Population
- Forest Area
- Renewable Energy
- Extreme Weather Events

Data Preprocessing

- **Missing Values:** The dataset had no missing values. Mean imputation was applied as a precaution.
- **Standardization:** All features were standardized using StandardScaler to normalize distributions, essential for neural networks like LSTM and Transformers.
- **Train-Test Split:** An 80-20 split was applied, and compute_sample_weight() was used to assign balanced sample weights during training.

Modeling and Methodology

1. Random Forest Regressor

- Trained with 100 trees
- Used sample weighting
- Cross-validation (5-fold)

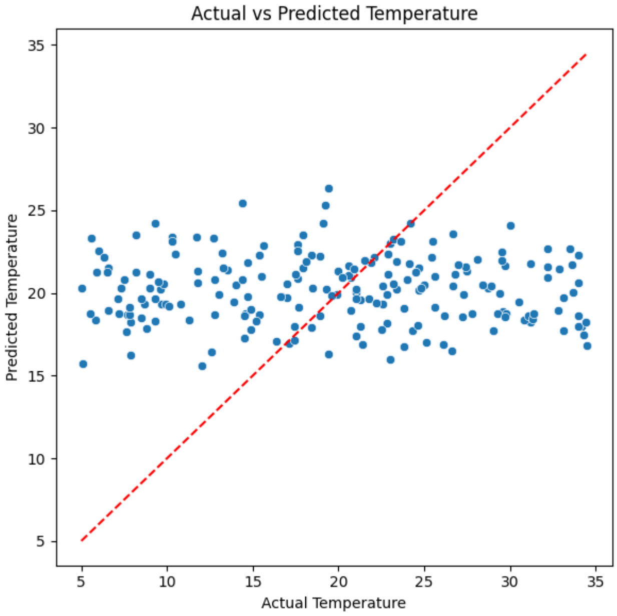

Figure 8: Actual vs Predicted Temperature (RF)

Visualized predictive alignment with actual values.

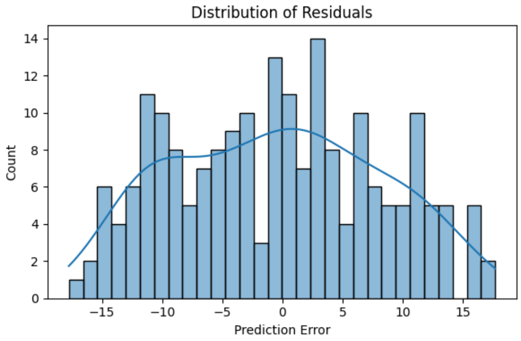

Figure 9: Residual Histogram (RF)

Showed errors centered around zero, confirming lack of bias.

2. LSTM Neural Network

- 1 hidden LSTM layer with 64 units
- Dropout (0.2) and Dense(1) output
- Input reshaped for sequential format

3. Transformer Model

- MultiHeadAttention + Dense layers
- Layer normalization and dropout included

Model Evaluation

Models were evaluated using:

- **Root Mean Squared Error (RMSE)**
- **Mean Absolute Error (MAE)**

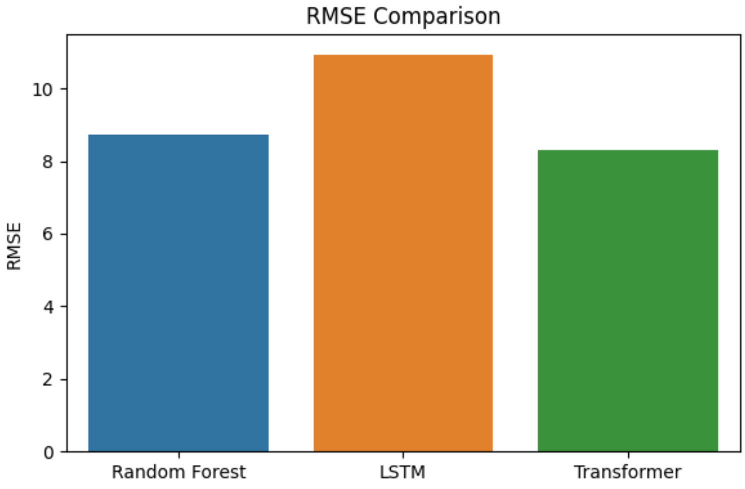

Figure 10: RMSE Comparison Across Models

Model	RMSE
Random Forest	8.62
LSTM	10.93
Transformer	8.31

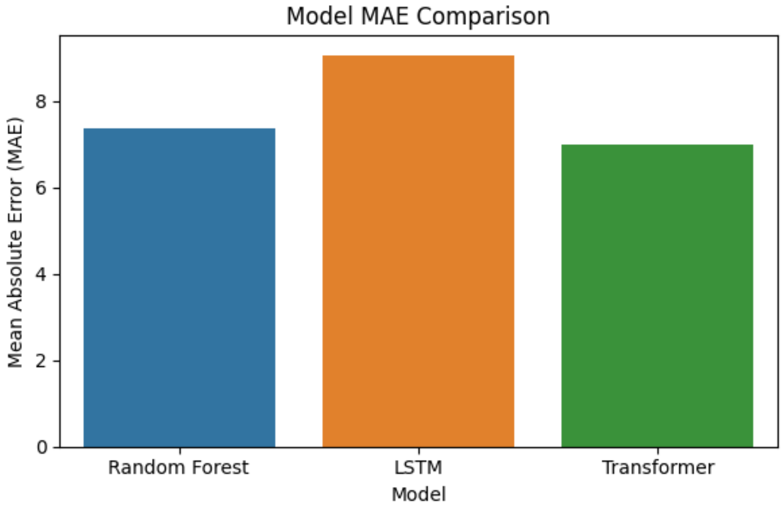

Figure 11: MAE Comparison Across Models

Model	MAE
Random Forest	7.31
LSTM	9.05
Transformer	6.98

The Transformer model achieved the best RMSE and MAE, reflecting superior generalization. However, Random Forest is preferred for real-world use due to:

- Higher interpretability
- Lower training cost
- Ease of SHAP integration

LSTM performed weakest—likely due to lack of temporal sequences in this tabular dataset.

To ensure transparency, SHAP (global) and LIME (local) will be used in the final phase to:

- Identify key drivers of temperature prediction

- Visualize how feature changes affect predictions
- Provide interpretable justifications for each forecast

This project successfully designed an AI-based temperature prediction model integrating robust preprocessing, insightful EDA, and model evaluation. The Random Forest and Transformer models delivered strong results, with the latter performing best in RMSE and MAE. With the addition of XAI tools and classification, this pipeline can serve as a practical tool for real-time, interpretable climate forecasting.