**Project Report: Carbon Footprint Forecasting using Machine Learning**

## **1. Introduction**

### **1.1 Overview**

Climate change has emerged as one of the most pressing global issues of the 21st century. A significant driver of global warming is the release of greenhouse gases, especially carbon dioxide (CO2), stemming from anthropogenic activities such as fossil fuel combustion and industrial processes. India, being a rapidly developing nation, faces the dual challenge of sustaining economic growth while mitigating environmental degradation. This research aims to forecast India's carbon emissions using machine learning approaches, thus supporting data-driven environmental policymaking.

### **1.2 Objective**

The primary objective of this project is to apply supervised learning techniques to historical emission data and build predictive models capable of estimating future CO2 emissions in India. This is expected to:

* Quantify the contribution of various industrial sectors to carbon emissions.
* Explore correlations among different emission sources.
* Identify the most impactful features contributing to fossil fuel-based emissions.
* Provide a prototype dashboard for real-time prediction.

### **1.3 Significance**

Accurate forecasting of emissions plays a vital role in strategic climate planning. It enables stakeholders and policymakers to design interventions tailored to high-emission sectors. This project not only enhances transparency in emission accounting but also introduces a scalable modeling pipeline for other developing nations.

## **2. Literature Review**

Past research has demonstrated the utility of data-driven approaches in climate forecasting. Statistical models like ARIMA and exponential smoothing have been used to model temporal emission trends. However, these models struggle with non-linear relationships between industrial activity and emissions.

Machine learning models, on the other hand, such as Random Forests and Gradient Boosting, have shown promise in:

* Capturing complex non-linear patterns
* Handling high-dimensional data
* Providing feature importance for interpretability

Hybrid models combining time series decomposition with supervised learning have also shown improved performance. Nonetheless, much of the literature focuses on global datasets, with relatively little focus on country-specific emission trends like those of India.

## **3. Methodology**

### **3.1 Data Collection**

* Source: Data was aggregated from public sources including Statista, official government environmental records, and industry whitepapers.
* Period: 1960 to 2023
* Features: Emissions by sector (power, transport, buildings), fuel-type (coal, oil, gas), and industrial processes (cement manufacturing, industrial combustion, etc.).

### **3.2 Data Preprocessing**

* Replaced missing data marked with '-' using NaN, followed by imputation using SimpleImputer with mean strategy.
* Removed non-numeric and timestamp formatting errors.
* Standardized features for regression modeling using StandardScaler.

### **3.3 Feature Engineering**

* Conducted correlation analysis to identify multicollinearity.
* Removed highly correlated or low-variance features.
* Created derived features to capture sector interactions (e.g., industrial combustion intensity).

## **4. Model Development**

### **4.1 Regression Models Applied**

* **Linear Regression**: Baseline model; interpretable but limited in capturing non-linearity.
* **Decision Tree Regression**: Handles non-linear dependencies; prone to overfitting.
* **Random Forest Regression**: Ensemble learning technique providing robustness and feature ranking.

### **4.2 Hyperparameter Optimization**

Used GridSearchCV for tuning key parameters:

* Tree depth
* Number of estimators
* Minimum samples split
* Cross-validation (k=5) to ensure generalization

### **4.3 Evaluation Metrics**

* **Mean Absolute Error (MAE)**: Measures average absolute errors.
* **Mean Squared Error (MSE)**: Penalizes larger errors more heavily.
* **R^2 Score**: Variance explained by the model.

## **5. Results and Interpretation**

| **Model** | **MAE** | **MSE** | **R^2 Score** |
| --- | --- | --- | --- |
| Linear Regression | ~7.12 | ~88.42 | 0.85 |
| Decision Tree | ~5.62 | ~72.30 | 0.89 |
| Random Forest | **3.84** | **51.79** | **0.93** |

* **Random Forest** outperformed other models in all metrics.
* Feature importance plot revealed:
  + Coal combustion and transportation are top predictors.
  + Power industry emissions significantly impact fossil fuel-based emissions.
  + Emissions from buildings and agriculture had lower predictive strength.

## **6. Visualization and Analysis**

* **Heatmaps** showed strong inter-feature correlations, aiding feature reduction.
* **Countplots and histograms** provided insights into distribution skewness and variance.
* **Feature importance bar plots** clarified sectoral priorities.
* **Prediction vs Actual scatter plots** validated model efficacy visually.

## **7. System Deployment and Use Case**

### **7.1 Dashboard Design**

A Flask-based web dashboard is under development with the following functionalities:

* Input: User-defined values for key emission sectors.
* Output: Real-time emission predictions.
* Visualizations: Sector-wise impact plots, temporal trends, policy simulation tools (planned).

### **7.2 Real-world Applications**

* Government agencies for emission control forecasting
* Industries to self-assess carbon compliance
* Researchers for climate trend modeling