EnviroMeter - Carbon Footprint Predictor

Dishita Shah – 16010122167 || Aakanksh Sen - 16010122166
Krish Satra – 16010122164 || Heemit Shah - 16010122168

Guided By:- Dr. Shila Jawale

SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Somaiya TRUST

Introduction

The world is facing a serious climate challenge, with global emissions crossing 37 billion tonnes of CO_2 each year. The average person contributes around 4.7 tonnes, yet most people don't realize how their daily habits travel, food, electricity, and consumption add to this impact.

Many individuals want to live more sustainably, but calculating a carbon footprint is confusing, time consuming, and often inaccurate because existing tools don't reflect real lifestyle patterns.

This creates a need for a simple and personalized solution. EnviroMeter fulfills this by turning everyday lifestyle inputs into clear carbon insights and practical steps to help users reduce their overall footprint.

Motivation

- Climate change is one of the most urgent global challenges, yet most individuals are unaware of how their daily lifestyle choices contribute to carbon emissions.
- People struggle to estimate their personal environmental impact because carbon footprint calculations are complex and depend on many factors like transport, diet, energy usage, and consumption habits.
- Existing tools are either too generalized or require domain expertise, making them inaccessible for regular users or students.
- There is a need for a simple, intelligent, data-driven system that can accurately estimate a person's carbon footprint and give personalized recommendations for reducing it.
- With the rise of machine learning, we can analyze real lifestyle data and build a smart predictor that helps individuals make environmentally responsible decisions.
- This project aims to empower users with awareness, make sustainability measurable, and encourage small changes that lead to a significant positive impact on the planet.

Problem Statement

The Problem:

- Individuals lack awareness of how their daily activities such as travel, food consumption, and energy usage contribute to their overall carbon footprint.
- Existing carbon footprint calculators are often complex, time-consuming, and provide generalized estimates that are not tailored to personal lifestyle patterns.
- There is no simple, accessible tool that offers personalized and actionable insights to help users reduce their environmental impact.

Our Solution:

EnviroMeter: An carbon footprint predictor that analyzes user behavior and delivers personalized carbon footprint assessments along with practical, data-driven recommendations for sustainable living.

Objectives

The objectives of our project are as follows:

1. To develop a model that accurately predicts an individual's carbon footprint based on lifestyle inputs.
2. To provide users with personalized, easy-to-understand insights about their environmental impact.
3. To recommend actionable steps and sustainable habits that help users reduce their carbon footprint.
4. To create a user-friendly interface that allows seamless data entry and real-time footprint calculation.
5. To promote environmental awareness by visualizing carbon emissions across different lifestyle categories (travel, energy, food etc.).

Scope

User Types:

- Individuals/Consumers: Enter lifestyle, travel, food, and household activity data to calculate their carbon footprint.
- Environmental Enthusiasts/Researchers: View aggregated insights and trends for awareness and analysis.

Platform Capabilities:

- A carbon footprint prediction using machine learning models trained on emission datasets.
- Personalized sustainability recommendations based on user habits and predicted footprint.
- Visualization dashboards showing category wise emissions (transport, energy, food, waste).

Technology Integration Scope:

- Machine Learning (XGBoost, Scikit-Learn) for carbon emission prediction.
- Flask API for backend model deployment and communication with the UI.
- React.js frontend for interactive user experience and data visualization.

End Output:

- A web based application that allows users to calculate their carbon footprint, understand key emission contributors, receive personalized reduction strategies, and visualize their environmental impact in a simple and meaningful way.

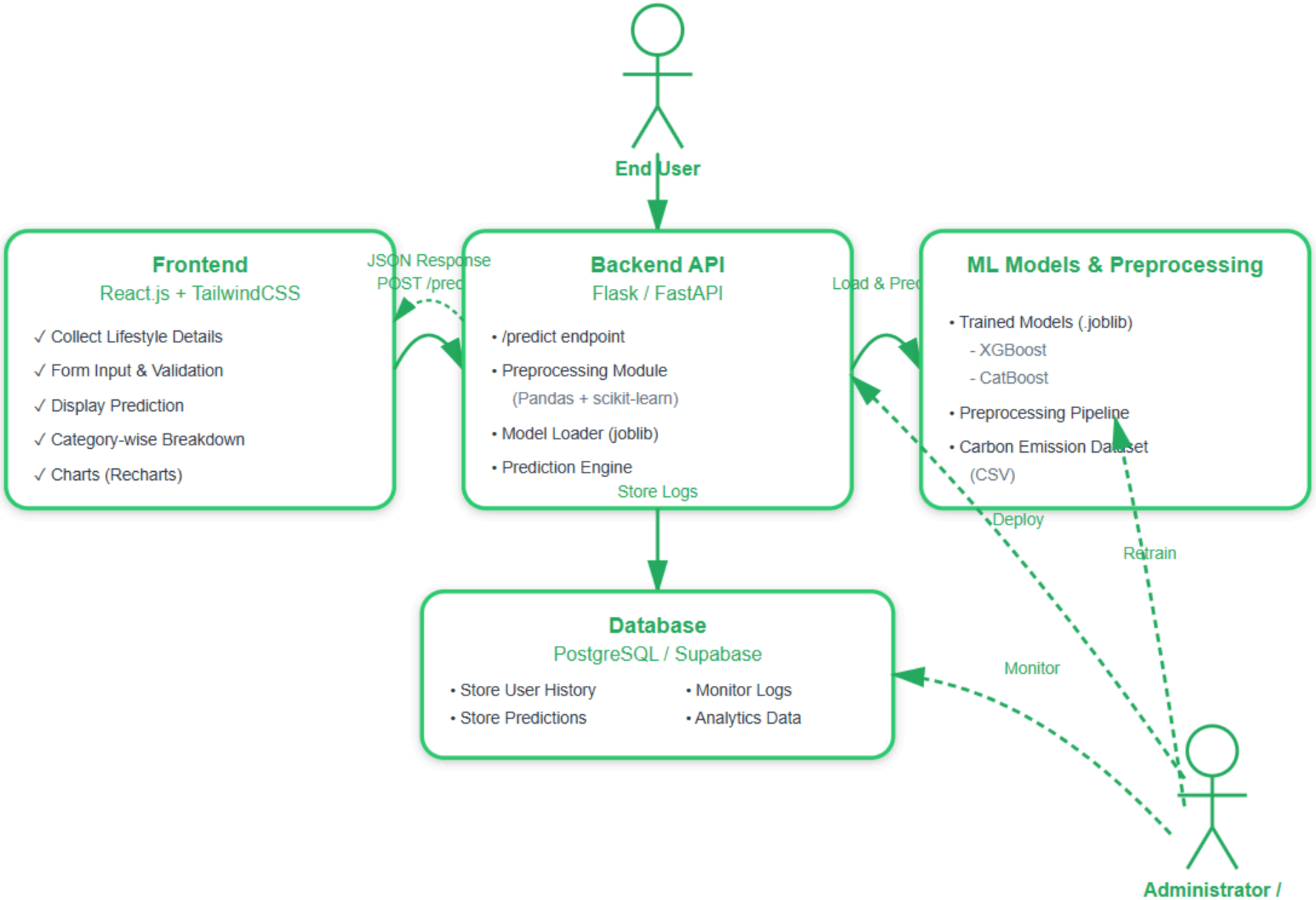
Comparison

Title of the Paper	Year of Publication	Methodology Used	Advantages	Limitations	Dataset Used	Results
Analysis & Prediction of Individual Carbon Footprints	2025	Linear Regression, RF, SVR, XGBoost, CatBoost	Web-integrated; highly accurate; identifies key contributors	Survey-based lifestyle inputs	Multi-factor lifestyle dataset	CatBoost $R^2 \approx 0.991$
Application of Carbon Footprint Clustering for Thrifty Food Plan Optimization	2024	DBSCAN Clustering + Optimization	Reduces emissions significantly	Complex model; database dependency	DataFRIENDS + NHANES + USDA	Reduced CO_2 from 106.2 → 94.2 kg/week
Carbon Prognosticator – Triple Ensemble Regressor & SHAP	2024	RF + CatBoost + DNN + SHAP	Very high accuracy; interpretable	High computation cost	Canadian vehicle + synthetic carbon dataset	$R^2 > 0.98$
Predictive Analytics for Carbon Footprint from Students' Activities	2023	SVR + Emission Factor Modeling	High R^2 (0.98); behavior-driven insights	Needs real-time student activity data	Student activity + electricity + commute dataset	MAE=129, RMSE=158, R^2=0.98
Household Carbon Footprint Estimation Using Machine Learning	2021	Random Forest, Gradient Boosting, Regression	High accuracy; identifies key contributors	Self-reported lifestyle data bias	Household lifestyle & energy survey	$R^2 \approx 0.87$
Activity-Based Carbon Footprint Calculation Using Behavioral Data	2020	Emission-factor mapping + ML	Uses IPCC emission factors; reliable	Requires large-scale surveys	National lifestyle survey	Accurate per-user emission calculations
Explainable ML for Carbon Emission Prediction	2022	XGBoost + SHAP	High interpretability	SHAP slow for large datasets	Energy + lifestyle dataset	MAE < 0.15 tons CO_2e
Energy Consumption & CO_2 Forecasting	2023	LightGBM, XGBoost, Ensembles	Handles mixed data well	Not individual-specific	UK/US energy datasets	XGBoost lowest RMSE
Carbon Footprint Modelling with ANN	2021	ANN (feedforward), Backprop	Good for complex patterns	Low interpretability	Household carbon dataset	RMSE < 0.20 tons CO_2e

Proposed tools and Methodologies

Module	Tools / Technologies / Models
Frontend	Next.js, TailwindCSS, Recharts, React-Hook-Form, Framer Motion
Backend	FlaskAPI (Python)
Machine Learning Model	CatBoost Regressor, Scikit-learn Pipeline
Data Processing	Pandas, NumPy, One-Hot Encoding, Feature Engineering

System Architecture



Dataset Description

Dataset Overview

- **Source:** Kaggle – *Individual Carbon Footprint Calculation Dataset*
- **Type:** Synthetic dataset generated from aggregated studies on lifestyle, transport, and energy habits.
- **Purpose:** To estimate total personal carbon emissions (in kg CO_2e) based on behavioral and lifestyle attributes.
- **Total Features:** 19 independent variables + 1 target variable (CarbonEmission).
- **Nature of Data:**
  - *Mixed-type* — includes categorical (e.g., diet, transport, recycling) and numerical (e.g., grocery bill, distance) features.
  - *Representative, not real* — simulates realistic household behavior patterns using statistically weighted distributions.

  Key Features
- **Lifestyle Factors:** Diet, Body Type, Frequency of Shower, Social Activity, New Clothes Purchased, TV/Internet Usage.
- **Energy Use:** Heating Energy Source, Energy Efficiency Preference, Cooking Devices.
- **Transportation:** Transport Mode, Vehicle Type, Vehicle Distance per Month, Air Travel Frequency.
- **Waste Generation:** Waste Bag Size & Weekly Count, Recycling Habits.
- **Expenditure Indicators:** Monthly Grocery Bill.
- **Target Variable:** CarbonEmission → total personal CO_2 equivalent output per individual.

Preprocessing Summary

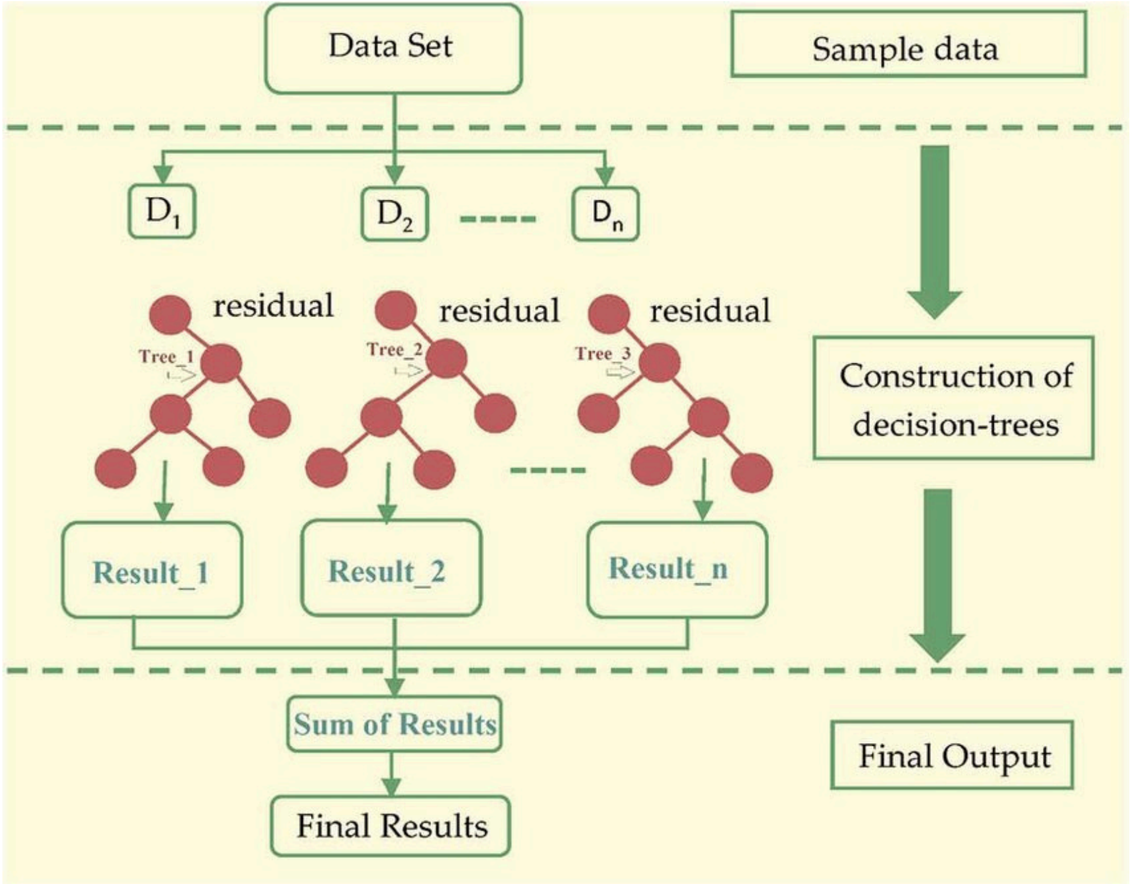
Before model training, data underwent systematic **feature engineering and transformation** steps to ensure compatibility and consistency:

- **Missing Value Handling**
  - Checked and imputed missing or inconsistent values.
  - Verified feature completeness since data was synthetically generated.
- **Feature Categorization**
  - Split all columns into **numerical** (e.g., distance, bills, usage hours) and **categorical** (e.g., transport, diet, recycling).
- **Scaling of Numerical Features**
  - Applied **standard normalization** to make features comparable and stabilize gradient-based algorithms.
- **Encoding of Categorical Features**
  - Converted textual categories (e.g., “Diesel”, “Electric”, “Vegetarian”) into **machine-readable vectors** using **One-Hot Encoding**.
- **Unified Transformation Pipeline**
  - Integrated both numeric and categorical transformations via a **ColumnTransformer**, ensuring a single, consistent preprocessing step for the full dataset.
- **Final Dataset Shape**
  - Approximately **10,000 samples × ~60 transformed features** after encoding and scaling.

XGBoost — Model Architecture & Hyperparameters

XGBoost architecture (tree boosting ensemble):

- Ensemble of gradient-boosted decision trees. Each tree fits residuals of the previous ensemble.
- Handles heterogeneous features, missingness, and interactions automatically.



XGBoost — Training & Evaluation Pipeline

Training Configuration

- **Boosting Rounds:** 300 trees (balanced between learning stability and training time)
- **Learning Rate:** 0.05 → slower, more stable convergence
- **Maximum Depth:** 6 → controls model complexity and avoids overfitting
- **Subsampling:** 0.8 → uses random samples per tree for better generalization
- **Column Sampling:** 0.8 → prevents dominance of specific features
- **Evaluation Metric:** Root Mean Squared Error (RMSE)
- **Objective Function:** Minimize squared error (reg:squarederror)
- **Early Stopping:** Stops if validation RMSE doesn't improve after 30 rounds

Training Process

- **Data Split:** Dataset divided into 80% training and 20% validation sets.
- **Model Fitting:** XGBoost trained iteratively, learning residual patterns between predicted and actual emissions.
- **Validation Monitoring:** Performance monitored on validation data to detect overfitting early.
- **Feature Importance Extraction:** After training, the most influential lifestyle features were analyzed.

Evaluation Strategy

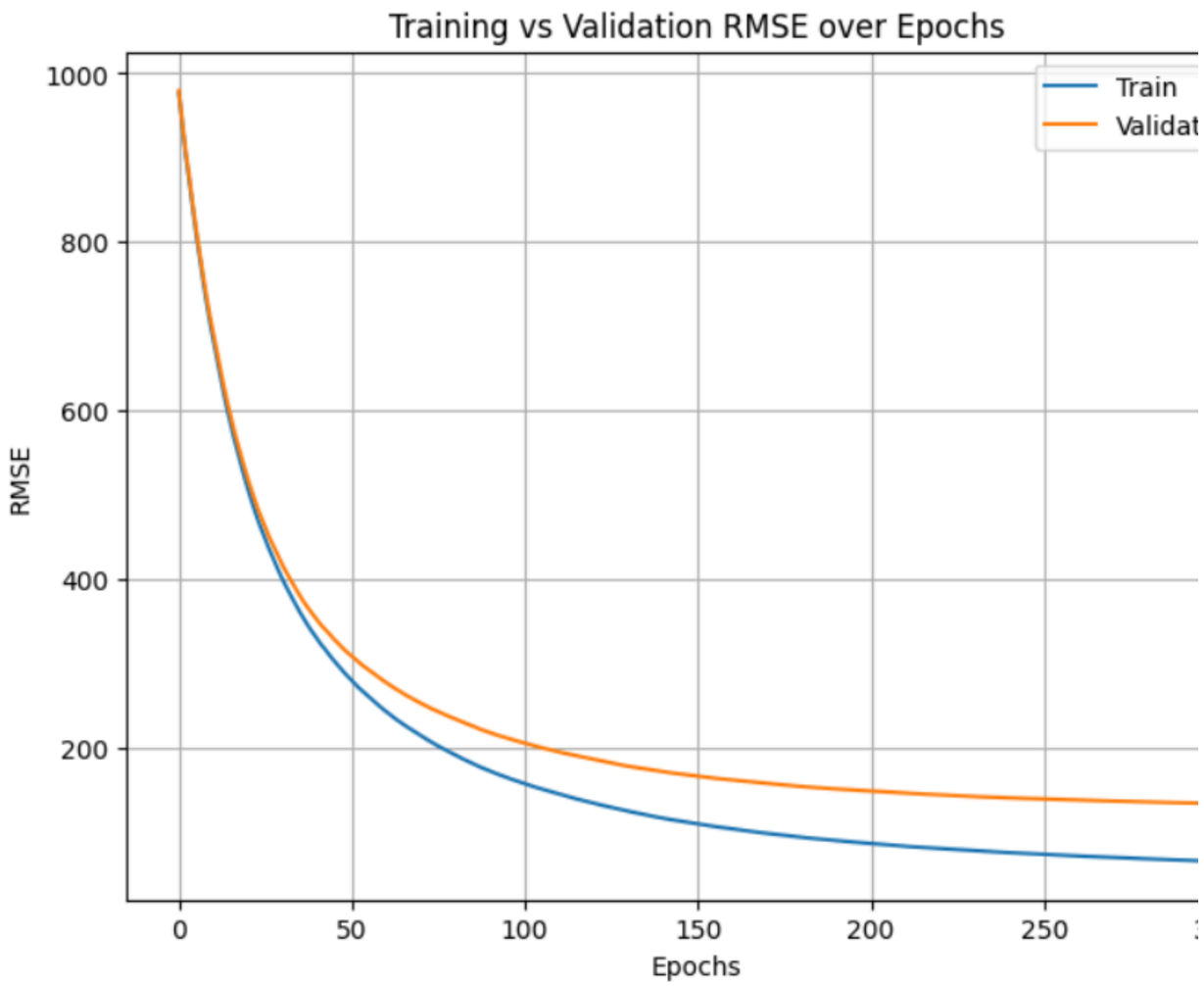
- **Primary Metrics:**
  - **MAE (Mean Absolute Error):** Measures average prediction error.
  - **RMSE (Root Mean Squared Error):** Penalizes large deviations more heavily.
  - **R^2 Score:** Represents how well the model explains variance in carbon emissions.
- **Cross-Validation:**
  - 5-fold validation confirmed model consistency across splits.
  - Average Cross-Validation $R^2 \approx$ **0.983 ± 0.001**

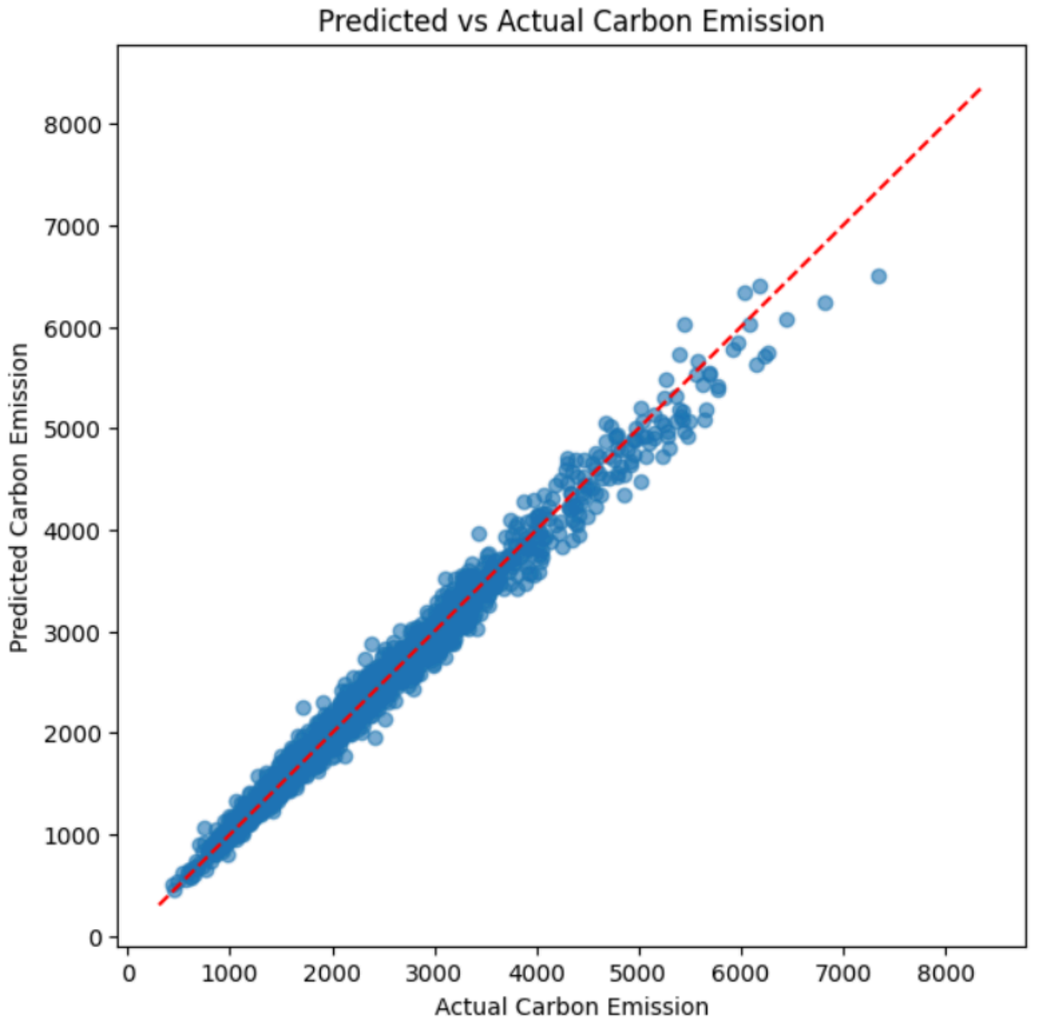
XGBoost — Results & Graphs

Model performance metrics (XGBoost):

- MAE: 94.621
- RMSE: 129.487
- R^2: 0.984

- Cross-Validation R^2: 0.983 ± 0.000
  Interpretation:
- The high R^2 indicates strong predictive capability; residual analysis should be inspected to confirm homoscedasticity.
- MAE and RMSE are in the units of the dataset (CO2e units) and show typical prediction errors - useful to map to real-world kgCO2e impacts.

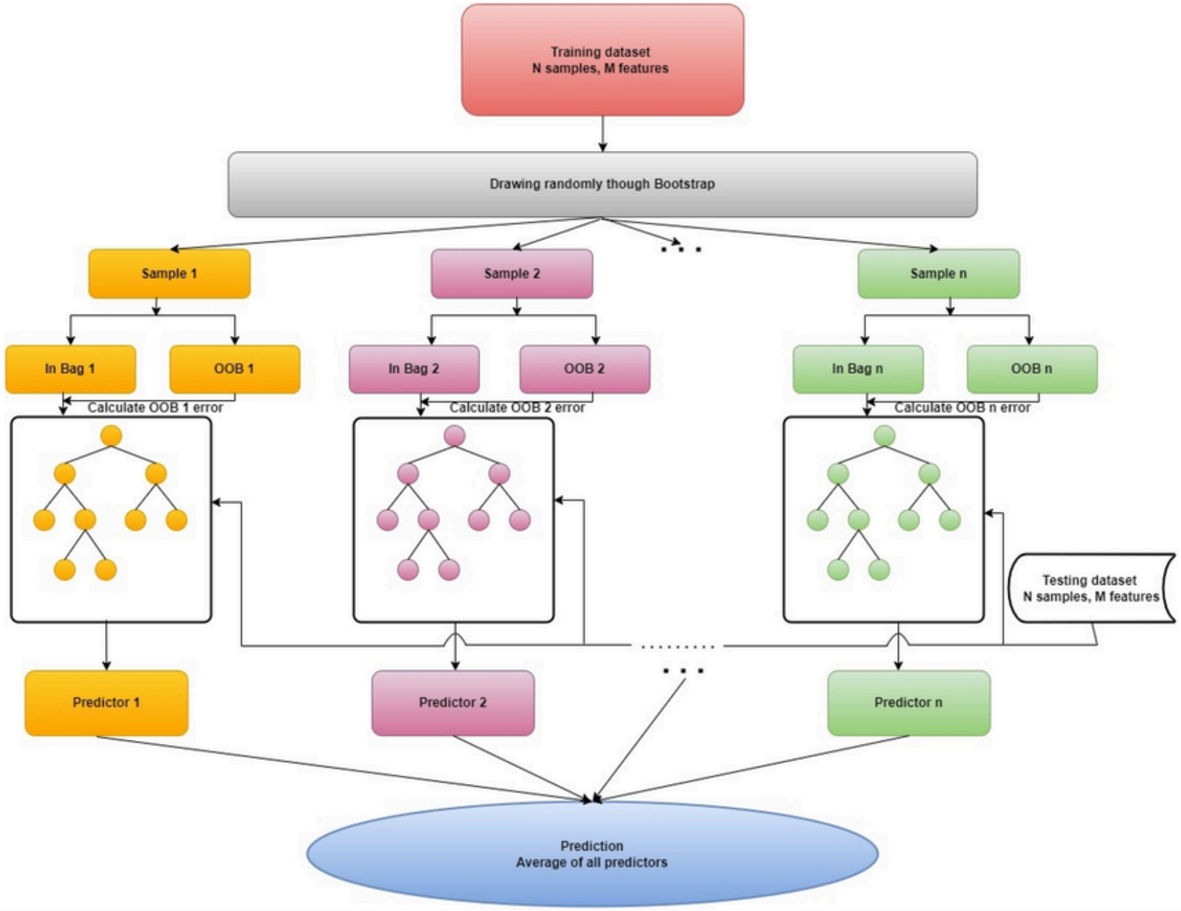

Training vs Validation RMSE over Epochs
Train
Validat
RMSE
Epochs
1000
800
600
400
200
0
50
100
150
200
250


Predicted vs Actual Carbon Emission
Predicted Carbon Emission
Actual Carbon Emission
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000
7000
8000

CatBoost — Model Architecture & Hyperparameters

CatBoost architecture (tree boosting ensemble):

- Gradient boosting on decision trees with ordered boosting to reduce target leakage in categorical encodings.
- Native categorical feature handling reduces need for one-hot encoding and preserves ordinal information where present.



CatBoost — Training & Evaluation Pipeline

Training Configuration

- **Iterations:** 500 boosting rounds (trees) to ensure thorough convergence.
- **Learning Rate:** 0.05 → balances learning speed and stability.
- **Tree Depth:** 8 → allows moderate model complexity to capture nonlinear interactions.
- **Loss Function:** Root Mean Squared Error (RMSE) → focuses on minimizing squared deviations.
- **Evaluation Metric:** RMSE on validation data to monitor overfitting.
- **Random Seed:** 42 for reproducibility.
- **Early Stopping:** Training stops automatically if validation RMSE does not improve for 30 consecutive rounds.
- **Verbose Logging:** Progress printed every 50 iterations for transparency.

Training Process

- **Data Split:** 80% training, 20% validation to ensure unbiased evaluation.
- **Native Categorical Handling:**
  - Categorical feature indices were passed directly to the model (cat_features), allowing **CatBoost** to internally encode categories efficiently using **ordered target statistics**.
- **Boosting Mechanism:**
  - Each new tree is built to correct the residuals from previous trees, with leaf values optimized using ordered boosting.
- **Regularization:**
  - Controlled through learning rate and early stopping to maintain generalization.

Evaluation Strategy

- **Performance Metrics:**
  - **MAE (Mean Absolute Error):** Measures the average deviation between predicted and actual values.
  - **RMSE (Root Mean Squared Error):** Highlights larger prediction errors more strongly.
  - **R^2 Score:** Indicates the proportion of variance in emissions explained by the model.
- **Validation:**
  - Model performance evaluated on a held-out validation set.
  - Stable convergence observed around 450 iterations, confirming robust fit without overfitting.

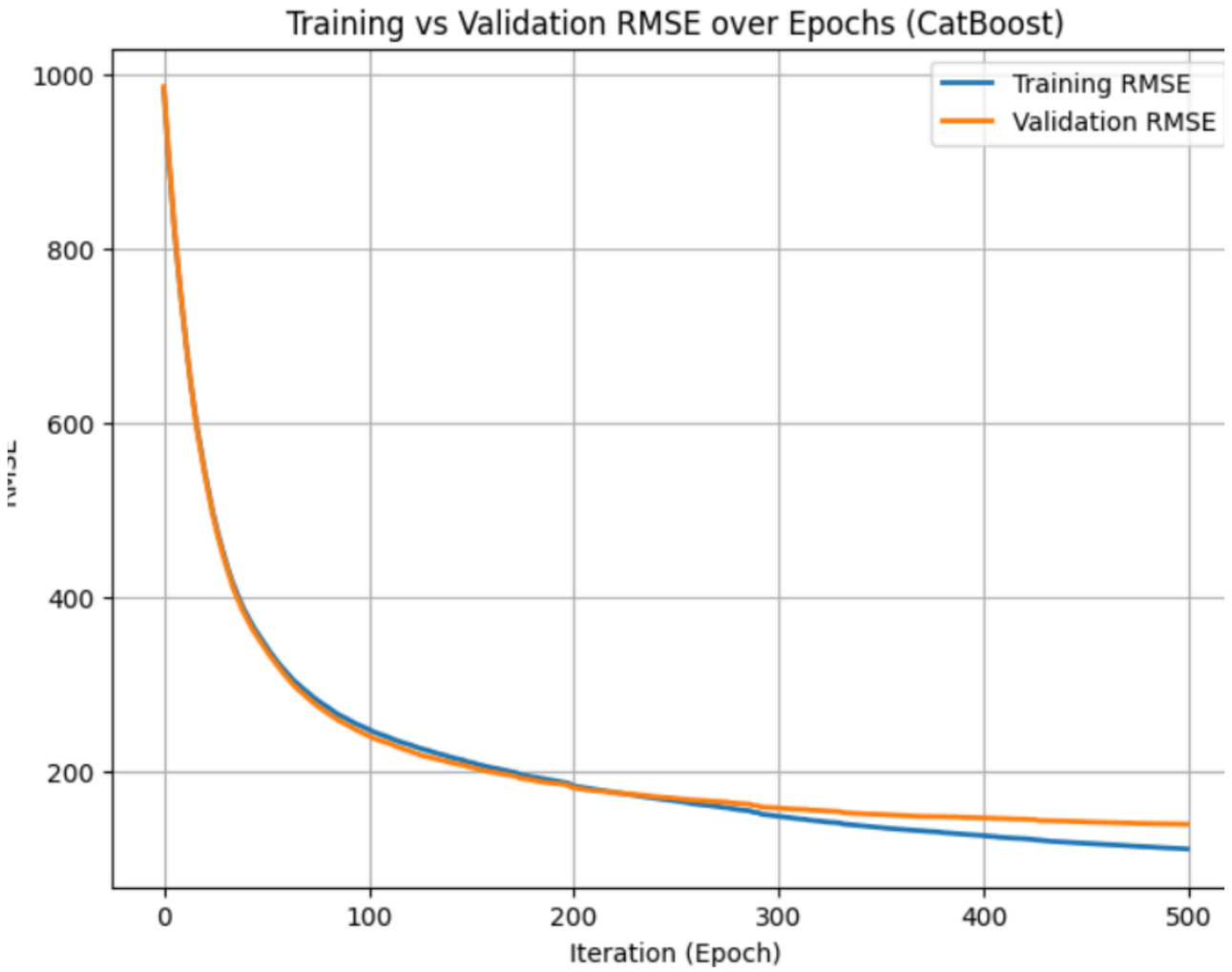
CatBoost — Results & Graphs

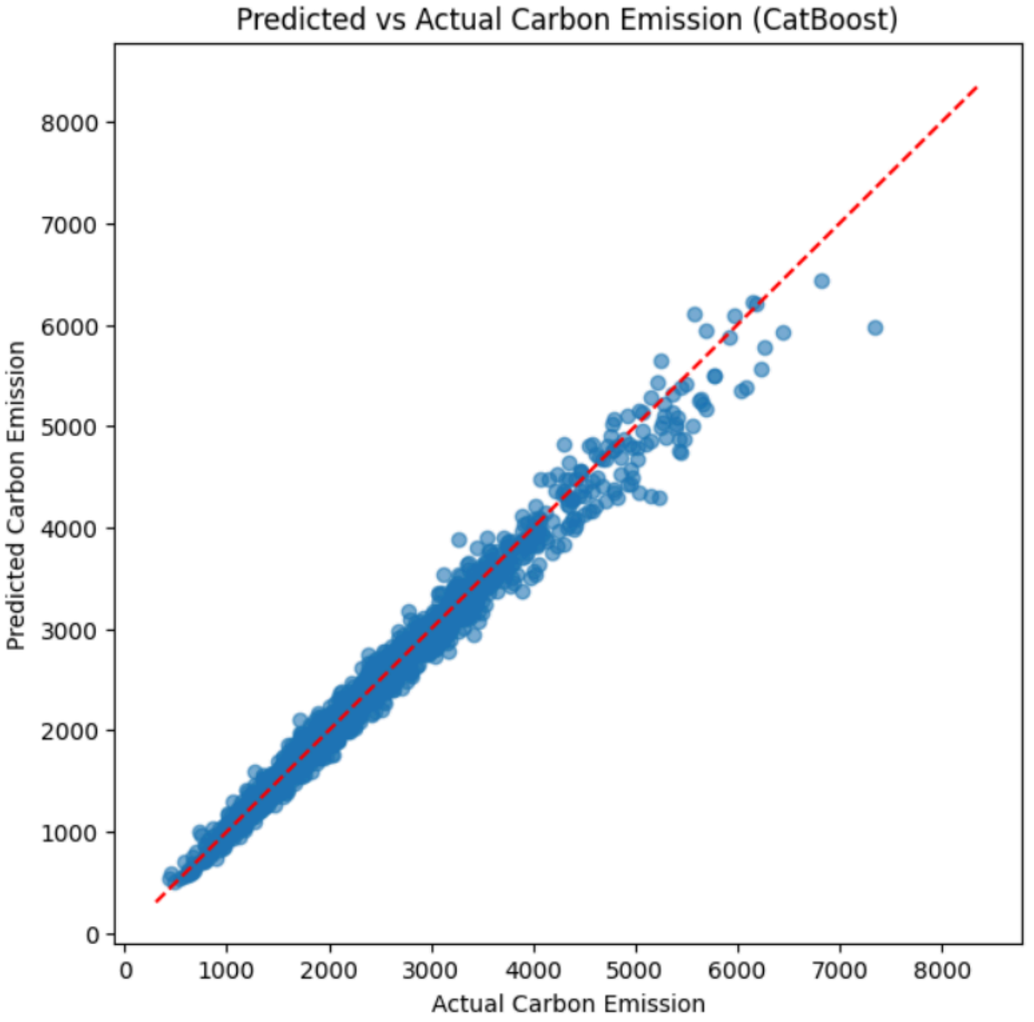
Model performance metrics (XGBoost):

- MAE: 96.445
- RMSE: 139.964
- R^2: 0.981

Interpretation:

- CatBoost produced competitive results with slightly higher RMSE but still high R^2 indicating reliable prediction.


Training vs Validation RMSE over Epochs (CatBoost)
Training RMSE
Validation RMSE
1000
800
600
400
200
0
100
200
300
400
500
Iteration (Epoch)


Predicted vs Actual Carbon Emission (CatBoost)
Predicted Carbon Emission
Actual Carbon Emission
0
1000
2000
3000
4000
5000
6000
7000
8000

SVM — Results & Graphs

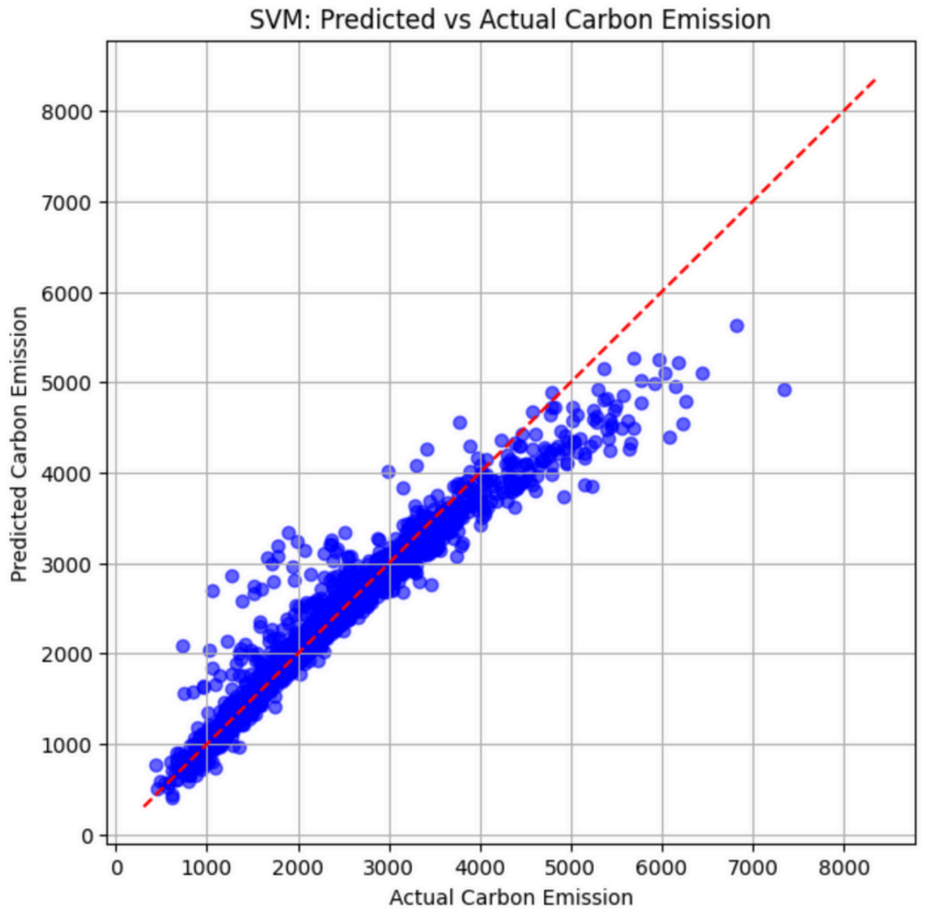
SVM MODEL PERFORMANCE

MAE: 153.441

RMSE: 274.472

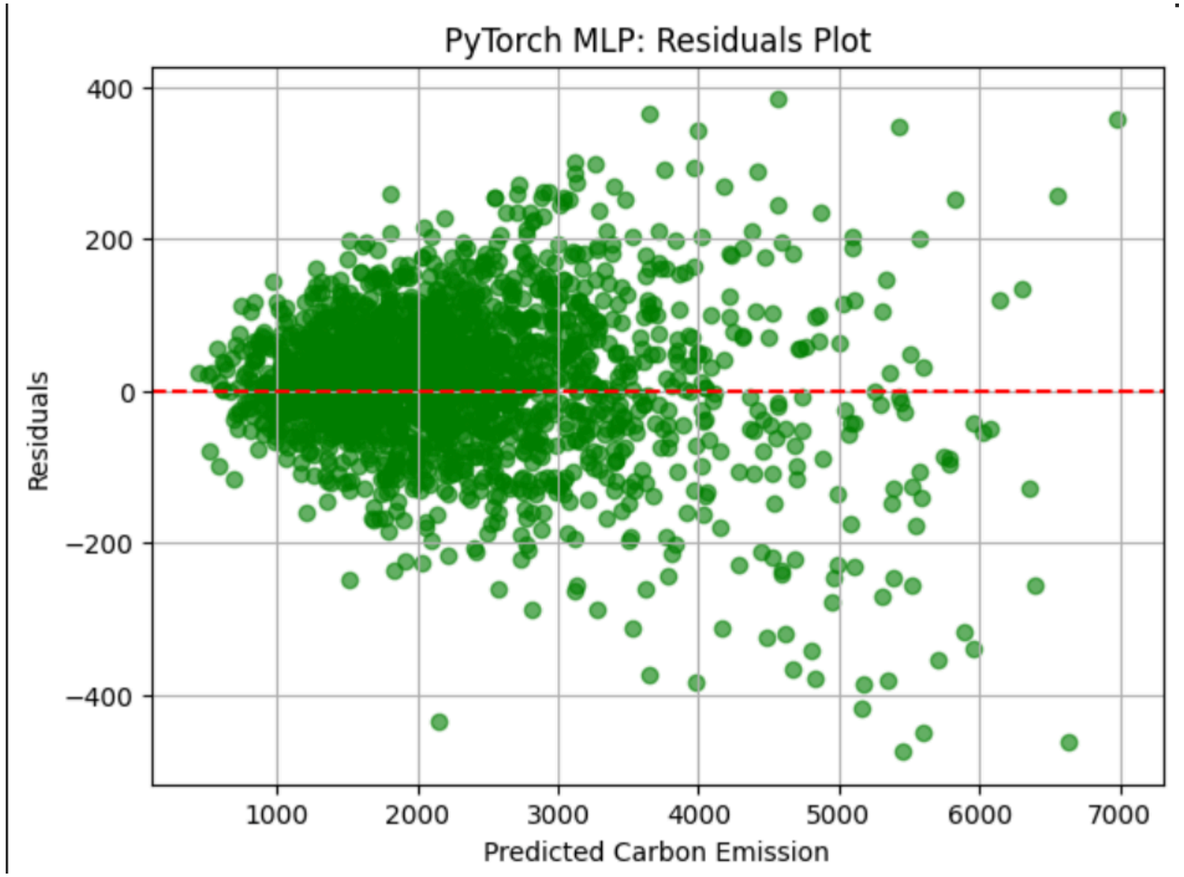
R^2: 0.928

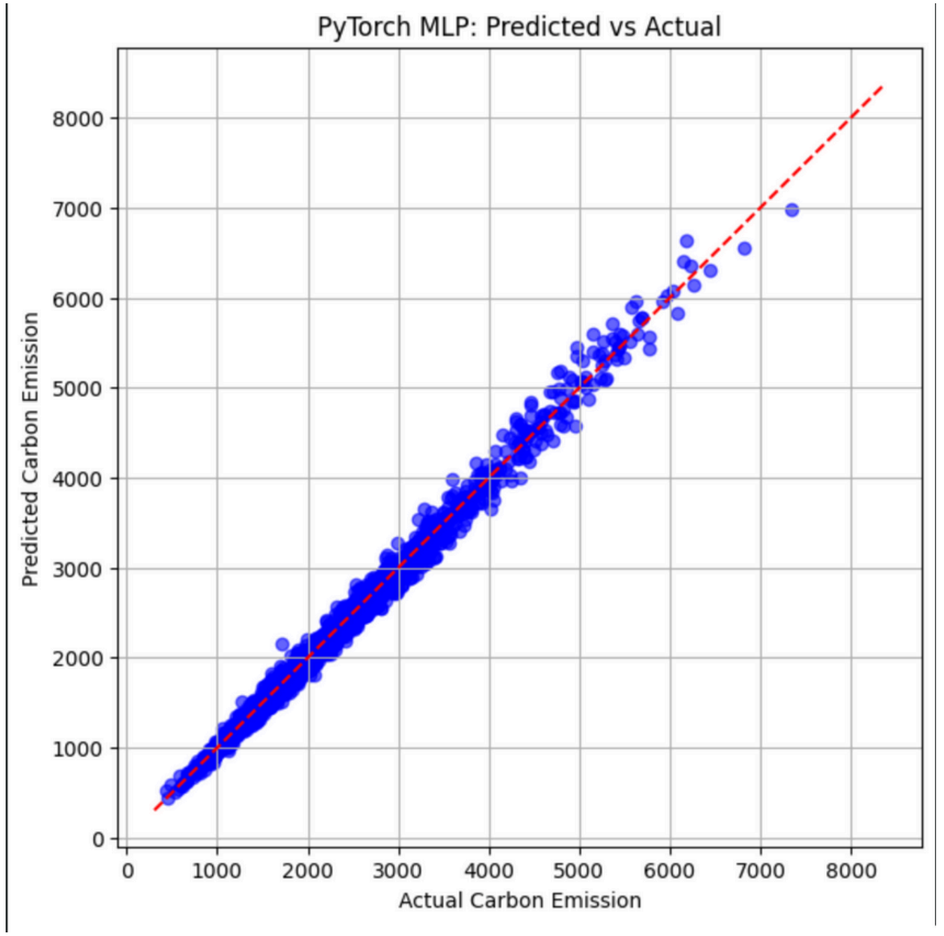

SVM: Residuals Plot
Residuals (Actual - Predicted)
Predicted Carbon Emission
2500
2000
1500
1000
500
0
−500
−1000
−1500
1000
2000
3000
4000
5000


SVM: Predicted vs Actual Carbon Emission
Predicted Carbon Emission
Actual Carbon Emission
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000
7000
8000

MLP — Results & Graphs

MLP MODEL PERFORMANCE
MAE: 72.959
RMSE: 98.776
R^2: 0.991


PyTorch MLP: Residuals Plot
Residuals
Predicted Carbon Emission
400
200
0
−200
−400
1000
2000
3000
4000
5000
6000
7000


PyTorch MLP: Predicted vs Actual
Predicted Carbon Emission
Actual Carbon Emission
0
1000
2000
3000
4000
5000
6000
7000
8000

SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Model Comparison Table

Model	MAE	RMSE	R^2	Remarks
SVM	153.441	274.472	0.928	Weakest performance; very high error; not suitable
MLP	72.959	98.776	0.991	Best accuracy; lowest error among all models
CatBoost	96.445	139.964	0.981	High accuracy; stable and reliable for deployment
XGBoost	94.621	129.487	0.984	Strong predictive power; low errors; highly dependable

Interpretation

- MLP achieved the best overall accuracy, but it is more complex, harder to tune, and less stable for mixed numerical + categorical data.
- XGBoost also performed strongly, but required more hyperparameter tuning and preprocessing steps.
- CatBoost offered the best balance of accuracy, stability, and deployment readiness, handling mixed feature types automatically while still achieving a high R^2 of 0.981.
- SVM performed the weakest, with high errors and lower reliability, making it unsuitable for this prediction task.

Future Scope

1. Advanced Model Explainability

Add SHAP-based explanations to show which lifestyle factors contribute most to a user's carbon footprint.

2. Personalized Reduction Recommendations

Generate tailored action plans and lifestyle changes to help users meaningfully reduce their emissions.

3. Simulation of Lifestyle Changes

Allow users to modify habits (diet, transport, energy use) and instantly see how their carbon emissions change.

4. Mobile App & API Integration

Extend the system into a mobile app and provide a public API so schools, NGOs, or other apps can integrate the prediction service.

Conclusion

The Carbon Emission Prediction System represents a modern, data-driven approach to environmental awareness and sustainability. By leveraging machine learning, behavioral data, and intuitive visualizations, it empowers users to understand the real impact of their daily lifestyle choices. The system not only delivers accurate emission estimates but also provides personalized, actionable recommendations that encourage meaningful change. With its scalable architecture, interactive design, and potential for real-world integration, this solution promotes responsibility, transparency, and long-term sustainability while advancing environmental consciousness among users.

Key Highlights:

- Predicts individual carbon emissions using advanced Machine Learning (XGBoost).
- Provides personalized, data-driven recommendations to reduce environmental impact.
- Simulates lifestyle changes to help users understand their CO_2 reduction potential.
- Offers a user-friendly interface with clear visualizations and explainability insights.
- Ensures accurate preprocessing and consistent prediction through a unified pipeline.
- Scalable design supports multi-user access, history tracking, and API integration.
- Promotes environmental awareness by translating complex data into simple, actionable insights.

References

Research Papers-

- A. Author, B. Author, “Analysis and Prediction of Individual Carbon Footprints: An Integrated Machine Learning and Web Application Approach,” IEEE, 2025.
- J. Chen, M. Li, and S. Gupta, “Application of Carbon Footprint Clustering for Thrifty Food Plan Optimization,” IEEE Big Data Conference, 2024.
- R. Singh, K. Verma, and L. Patel, “Carbon Prognosticator: A Triple Ensemble Regressor and SHAP Analysis for Enhanced CO_2 Emission Modeling,” Elsevier, 2024.
- T. Nguyen and P. Rodriguez, “Predictive Analytics Model for Optimizing Carbon Footprint from Students’ Learning Activities in Computer Science Majors,” IEEE Access, 2023.

- A. Author, B. Author, “Analysis and Prediction of Individual Carbon Footprints: An Integrated Machine Learning and Web Application Approach,” IEEE, 2025.
- J. Chen, M. Li, and S. Gupta, “Application of Carbon Footprint Clustering for Thrifty Food Plan Optimization,” IEEE Big Data Conference, 2024.
- R. Singh, K. Verma, and L. Patel, “Carbon Prognosticator: A Triple Ensemble Regressor and SHAP Analysis for Enhanced CO_2 Emission Modeling,” Elsevier, 2024.
- T. Nguyen and P. Rodriguez, “Predictive Analytics Model for Optimizing Carbon Footprint from Students’ Learning Activities in Computer Science Majors,” IEEE Access, 2023.

SOMAIYA
VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Thank You