ML Based Assessment of Household Carbon Emission in Nepal

Bipash Lamsal[1], Biraj Subedi[2,*], Jaydev Pandey[3], Hemantaraj Dhungana[4], Binod Sapkota[5],
[1,2,3,4]Himalaya College of Engineering,Tribhuvan University(TU), Nepal
[5]Institute of Engineering (IOE), Thapathali Campus, Tribhuvan University(TU), Nepal
*Corresponding author: hce077bct016@hcoe.edu.np

Abstract
This study develops a machine learning (ML)-based framework to assess and mitigate household carbon emissions in Nepal, leveraging a stacking ensemble model (Random Forest + Gradient Boosting meta-regressor) that achieves high predictive accuracy (MSE: 112.17, R^2: 0.98). By analyzing data from 4,000 households across energy use, transportation, waste, and dietary habits—collected via a structured Google Forms survey and processed using feature selection and Z-score normalization—the system provides personalized carbon footprints and reduction strategies, validated against IPCC benchmarks. The web-based FastAPI-React tool identifies high-impact factors (e.g., LPG consumption, bottled water usage, rainwater harvesting) and effective mitigation measures (e.g., solar adoption), offering actionable insights for households and policymakers to support Nepal's climate goals. This work advances scalable, context-aware ML solutions for sustainability in developing regions.

Keywords: Machine learning, Carbon footprint, Stacking ensemble, Household emissions, Climate mitigation

1. Introduction

Climate change poses significant challenges for Nepal, where rapid urbanization and traditional rural practices create unique emission patterns. Although Nepal's per capita emissions remain low, household-level contributions are increasing due to rising energy demand, transportation, and consumption. Existing carbon calculators—designed primarily for Western contexts—fail to capture these regional nuances, creating a critical gap in climate-action tools. Traditional carbon accounting methods rely on generalized emission factors and manual calculations, often lacking personalization and scalability. Recent advances in ML (Machine Learning) offer opportunities to develop data-driven solutions that account for local behaviors and infrastructure. However, applications in low-resource settings remain limited due to data scarcity and computational constraints. This study presents an ML-powered carbon footprint assessment system tailored for Nepalese households. Our approach combines:

- A stacking ensemble model for high-accuracy prediction.
- Context-specific emission factors (e.g., $0.9\ \mathrm{kgCO_{2e}/kWh}$ for Nepal's grid).
- A user-friendly web interface for accessibility.

The system contributes to Nepal's Nationally Determined Contributions (NDCs) by enabling evidence-based mitigation strategies at the household level.

2. Literature Review

Carbon footprint assessment has evolved from manual accounting toward data-driven and ML-based approaches. Early studies emphasized behavioral awareness and feedback mechanisms for emission reduction [1], while more recent work leverages ML to enhance estimation accuracy and scalability. For instance, [2] applied ML methods for urban carbon emission monitoring, demonstrating their potential for policy support and targeted interventions. At the household level, three methodological frameworks are commonly used:

- **Activity-based methods**: Estimate emissions from direct consumption data using IPCC emission factors [7].
- **Input–output analysis (IOA)**: Links household expenditure to sectoral emissions at the national or regional level [8].
- **Machine learning models**: Predict emissions based on behavioral and socioeconomic predictors [3], [4].

Recent ML-based applications have shown high predictive accuracy. [3] achieved $R^2 > 0.96$ using hybrid ensemble models, while [4] demonstrated gradient-boosting approaches for building-level emissions. Additional works have explored residential and household-level predictions in Asian contexts. For example, [12] showed that incorporating behavioral indicators, such as cooling habits and cooking frequency, improved emission prediction accuracy in Indian households. Similarly, [13] introduced a hybrid socio-technical ML framework for residential footprints, highlighting the importance of demographic and lifestyle variables. In the South Asian region, localized studies remain limited and often rely on secondary or aggregated data. [5] highlighted Nepal's lack of household-level datasets and localized emission factors as a key research barrier. [6] similarly emphasized the heterogeneity of household energy behavior across geography and income groups. Expanding this perspective, [11] analyzed household carbon emissions in South Asia and found significant variation driven by behavioral and infrastructural differences, supporting the need for region-specific modeling. Ensemble learning methods have increasingly been adopted for emission prediction. [14] demonstrated that stacking-based ensemble

Article History:
Submission: 2025-07-01, Revision: 2025-11-07, Acceptance: 2025-11-18

models outperform single-model approaches for household energy emissions, reinforcing the suitability of the stacking framework used in this study. Given these limitations and opportunities, the present study contributes by developing a machine learning–based estimation framework for household carbon emissions in Nepal. While the emission coefficients are based on internationally recognized IPCC and literature-derived factors, they are applied to Nepal's socioeconomic and energy-use data to produce more context-aware predictions. The proposed stacking ensemble model enhances predictive robustness, and the accompanying web-based tool provides an accessible interface for household-level carbon awareness and mitigation planning.

3. Objectives

The main objective of this study is to develop an accurate, region-specific, and user-friendly tool to estimate household carbon emissions in Nepal. Specifically, the goals are:

- To implement an interactive carbon footprint calculator using a stacking ensemble model for robust emission prediction.
- To provide actionable insights and personalized recommendations for emission reduction in areas such as energy, transportation, waste, and diet.
- To support climate awareness, policy formulation, and sustainable behavior through accessible digital technology.

4. Methodology

4.1 Data Collection

We conducted a structured survey using Google Forms between December 2024 and February 2025, receiving responses from over 4,000 households throughout Nepal. The questionnaire consisted of structured, close-ended questions using multiple-choice and 5-point Likert scales to quantify household practices (e.g., frequency of LPG use or meat consumption). Data were collected across all seven provinces, ensuring representation of both urban and rural households. After preprocessing and cleaning incomplete or inconsistent responses, the final dataset contained **4,089 households**. The survey collected data on:

- **Energy consumption:** Electricity, LPG, and firewood usage.
- **Transportation:** Private vehicle use, public transport, air travel.
- **Diet:** Frequency of meat, dairy, and packaged food consumption.
- **Water and waste:** Use of bottled water, rainwater harvesting, recycling practices.

4.2 Target Variable and Feature Selection

The target variable is the **total household carbon emissions** measured in kg CO_2e per month. Features include energy consumption, transportation, diet, and water/waste practices.

Feature selection was performed using:

- Correlation heatmaps to remove highly collinear features (Pearson correlation > 0.8)
- Domain knowledge to retain variables relevant to household carbon emissions
- Exploratory analysis to ensure variability and predictive relevance

4.3 Preprocessing and Feature Engineering

- **Categorical Encoding:** One-hot and label encoding techniques were applied
- **Normalization:** Z-score normalization standardized numeric values
- **Handling Missing Values:** Numeric missing values were imputed with the median; categorical with the mode
- **Outlier Treatment:** Values beyond 3 standard deviations were capped at the 99th percentile

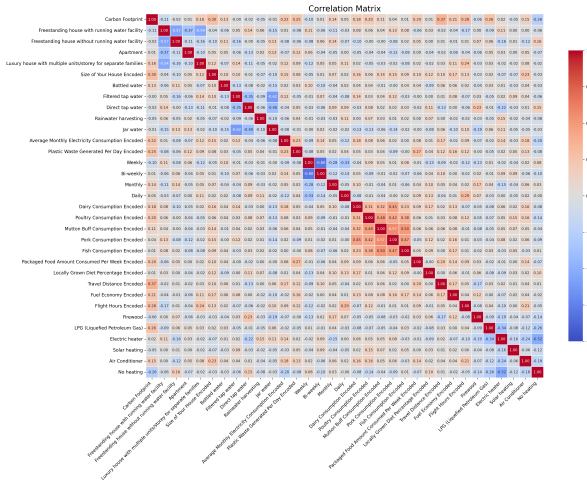

Figure 1. Correlation Heatmap of Features used in Modeling

4.4 Data Splitting

The dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling. A fixed random seed ensured reproducibility of results.

4.5 Emission Factor Mapping

To quantify carbon emissions, we applied emission factors sourced from the IPCC ipcc2019refinement and FAO fao2013livestock:

- **Electricity:** 0.9 kg CO_2e per kWh (based on Nepal's energy mix)
- **Meat consumption:** 27 kg CO_2e per kg
- **LPG:** 2.983 kg CO_2e per kg
- **Bottled water:** 3 kg CO_2e per liter

4.6 Model Architecture

We implemented a stacking ensemble model composed of:

- **Base Models:** Random Forest Regressor (RF), Gradient Boosting Regressor (GBR)
- **Meta-Model:** Multivariable Linear Regression (MLR)

Stacking was selected because it combines the strengths of multiple base models, reducing bias and variance, and often achieves higher predictive accuracy than any single model alone. This approach is particularly suitable for household carbon emission estimation, where the relationships between features and emissions are complex and non-linear.

This architecture is expressed mathematically as:

$$\hat{y} = \beta_0 + \beta_1 f_{RF}(x) + \beta_2 f_{GB}(x) \quad (1)$$

Where:

- $\hat{y}$ = Predicted household carbon emissions (kg CO_2e per month)
- $f_{RF}(x)$ = Prediction from Random Forest
- $f_{GB}(x)$ = Prediction from Gradient Boosting
- $\beta_0, \beta_1, \beta_2$ = Coefficients learned by the meta-regressor (MLR)

4.7 Model Implementation

The model was implemented in **Python 3.10** using:

- `pandas` and `NumPy` for data manipulation
- `scikit-learn` for modeling and evaluation
- `matplotlib` and `seaborn` for visualization

Stacking was implemented with `StackingRegressor` and 5-fold cross-validation.

4.8 Model Evaluation

Model performance was evaluated using:

- **Mean Squared Error (MSE):**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

Where: y_i = actual emissions, $\hat{y}_i$ = predicted emissions, n = total samples

- **R^2 Score:**

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (3)$$

Where: $\bar{y}$ = mean of actual emissions, measuring goodness of fit

4.9 Web Tool Integration

A web-based carbon calculator was developed:

- **Backend:** FastAPI serving model predictions via REST API
- **Frontend:** ReactJS interface for household input and visualization
- **Deployment:** Docker containers for reproducibility and scalability

5. Results and Discussion

5.1 Performance Comparison

Performance comparison for different Regression model is presented in Table 1.

Table 1. Performance of Regression Models

Model	MSE	R^2
Multivariable Linear Regression	4300	0.40
Random Forest Regressor	235.39	0.93
Gradient Boosting Regressor	597.34	0.89
Stacking Ensemble Regressor	**112.17**	**0.98**

5.2 Residual and Error Analysis

Residuals from the stacking model were tightly distributed around zero, indicating minimal bias and high predictive reliability. The error distribution showed slight left-skewness, suggesting rare overestimation but overall low variance.

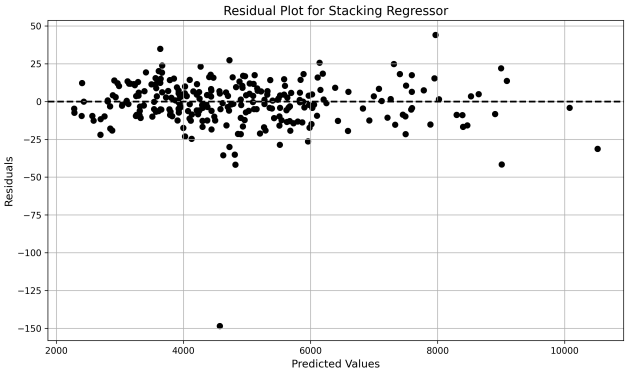

Figure 2. Residual Distribution for Stacking Model Predictions

5.3 Learning Behavior

The learning curve demonstrated that model performance stabilized after training on approximately 3,000 data points, indicating strong generalization and minimal overfitting.

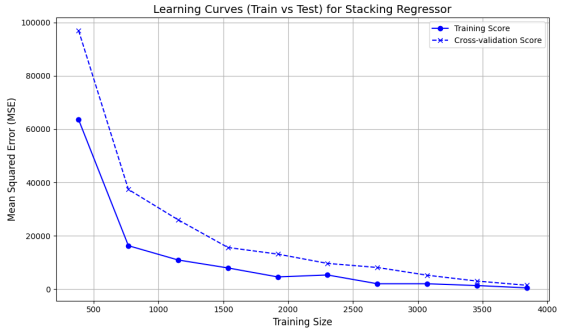

Figure 3. Learning Curve Indicating Convergence of the Stacking Model

5.4 Feature Impact Analysis

Permutation importance analysis identified the most influential features:

- LPG Consumption
- Electricity Usage
- Bottled Water Consumption
- Frequency of Meat Intake
- Presence of Rainwater Harvesting

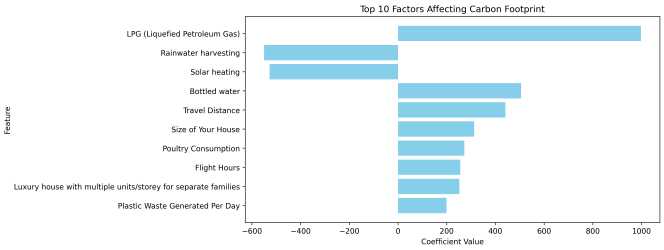

Figure 4. Feature Importance Scores from Permutation Analysis

6. Conclusion and Future Work

6.1 Conclusion

The ML-based assessment tool for household carbon emissions in Nepal successfully demonstrated the use of a stacking ensemble model (Random Forest + Gradient Boosting + MLR) to provide accurate carbon footprint predictions. The model achieved high performance with an MSE of 112.17 and an R^2 of 0.98, validating its robustness and generalization capacity. The tool provided personalized recommendations to users, encouraging more sustainable household practices.

Key influencers of carbon emissions—such as LPG usage, bottled water consumption, and the presence or absence of rainwater harvesting systems—were effectively identified. These insights can support both individual behavior change and policy interventions. The system aligns with Nepal's commitment to reduce emissions and strengthens efforts toward climate resilience.

6.2 Limitations

Despite its success, the project had some limitations:

- **Data Quality:** Self-reported survey data may include biases or inaccuracies, affecting prediction accuracy.
- **Lack of Real-time Input:** The model operates on static survey responses and does not incorporate live energy or consumption data.

6.3 Future Enhancements

Several improvements are proposed for future versions:

- **IoT Integration:** Real-time data from smart meters and connected devices can improve accuracy and allow dynamic tracking.
- **Mobile App Development:** A mobile interface will increase user reach and provide functionalities like push notifications, goal setting, and progress tracking.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] West, S., Smith, T., Patel, R. (2015). Evaluating the effectiveness of a carbon footprint calculator: The role of personalized feedback. *Environmental Research Letters*, 10(4), 045001.

[2] Gill, R., Tan, M., Choudhary, A. (2024). Monitoring urban emissions using machine learning: A case study approach. *Environmental Modelling & Software*, 170, 105728.

[3] Boateng, G. O., Du, Z., Qian, Y. (2020). Predicting carbon emissions with machine learning: A hybrid model approach. *Sustainable Cities and Society*, 53, 101938.

[4] Zheng, L., Wang, X., Liu, M. (2024). Predicting building carbon emissions using ensemble machine learning models. *Energy and Buildings*, 289, 113006.

[5] Sharma, R., Poudel, S., Ghimire, B. (2021). Analysis of greenhouse gas emissions and challenges in data availability: A case from Nepal. *Journal of Environmental Management*, 290, 112548.

[6] Bhattarai, A., Shrestha, M. (2022). Rural household energy use and carbon emission patterns in Nepal: A regional assessment. *Energy for Sustainable Development*, 68, 122–134.

[7] Hertwich, E. G., Peters, G. P. (2009). Carbon footprint of nations: A global, trade-linked analysis. *Environmental Science & Technology*, 43(16), 6414–6420.

[8] Lenzen, M., Moran, D., Kanemoto, K., Geschke, A. (2012). Mapping the structure of the world economy. *Environmental Science & Technology*, 46(15), 8374–8381.

[9] Intergovernmental Panel on Climate Change (IPCC). (2019). *2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories*. Retrieved from https://www.ipcc-nggip.iges.or.jp/public/2019rf/

[10] Gerber, P. J., Steinfeld, H., Henderson, B., Mottet, A., Opio, C., Dijkman, J., Falcucci, A., Tempio, G. (2013). *Tackling climate change through livestock: A global assessment of emissions and mitigation opportunities*. Food and Agriculture Organization of the United Nations (FAO). Retrieved from https://www.fao.org/3/i3437e/i3437e.pdf

[11] Paudel, R., Uddin, S. M. (2023). Household carbon emissions in South Asia: Infrastructure, lifestyle, and behavioral determinants. *Energy Policy*, 180, 113600.

[12] Kumar, A., Singh, R. (2022). Machine learning-based prediction of household electricity-related carbon emissions in India. *Sustainable Energy Technologies and Assessments*, 52, 102305.

[13] Wang, H., Li, Z., Chen, X. (2023). A socio-technical model for predicting residential carbon footprints using hybrid machine learning. *Journal of Cleaner Production*, 412, 137423.

[14] Li, Y., Zhao, P. (2024). Enhancing household carbon emission prediction using ensemble learning: A stacking-based approach. *Applied Energy*, 352, 121987.