A Machine Learning Model For Agricultural CO_2 Emissions Prediction

Sathwika.N,Pavan Kumar.V,Mani Vardhan.S , Dr.Debanjali Bhattacharya
School of Computing, Amrita Vishwa Vidyapeetham, Bengaluru, India
{bl.en.u4aie23119, bl.en.u4aie23134, bl.en.u4aie23158}@bl.students.amrita.edu, b_debanjali@blr.amrita.edu

***Abstract*—Agricultural activities, such as crop cultivation, livestock farming, and the use of chemical inputs, contribute significantly to global CO_2 emissions, affecting climate patterns and disrupting ecosystems. This study focuses on predicting CO_2 emissions from agricultural data using machine learning models, including XGBoost, CatBoost, Random Forest, MLPRegressor, Gradient Boosting, Ridge Regression and a Voting Regressor ensemble. These models are selected for their ability to capture complex relationships, considering factors like crop types, irrigation practices, soil quality, fertilization levels, and weather data. By evaluating their performance, the study aims to identify the most robust prediction approach, offering insights into agricultural CO_2 emissions. The goal is to develop strategies for reducing carbon footprints and promoting sustainable practices.**

***Index Terms*—CO_2 emissions, Machine Learning, XGBoost, CatBoost, Ensemble Learning, MLPRegressor, Voting Regressor, Agriculture, Prediction Models**

I. Introduction

The agricultural sector is a major emitter of greenhouse gases, especially CO. Accurate prediction of emissions from agricultural practices is crucial to formulate effective mitigation strategies. With advancements in machine learning, data-driven models offer powerful tools for environmental prediction tasks. This project leverages multiple machine learning algorithms to predict CO emissions from agricultural data, comparing their performance to recommend the most effective model. The study incorporates key agricultural parameters such as fertilizer usage, crop types, irrigation levels, and land area to train and evaluate the models. Models used include XGBoost, CatBoost, Random Forest, Gradient Boosting, Ridge Regression, MLP Regressor, and an ensemble Voting Regressor. The goal is to identify a model that provides high accuracy and generalizability, thereby supporting sustainable agriculture practices and aiding policymakers in climate-resilient planning.

II. Literature Survey

[1]Recent advancements in machine learning (ML) have revolutionized the prediction of carbon dioxide (CO) emissions in agricultural systems. In the study by Hassan et al. (2024), various hybridized ML models such as Additive Regression-Random Forest (AR-RF), Multi-Scheme Random Forest (MS-RF), and Iterative Classifier Optimizer integrated with AR-RF (ICO-AR-RF) were employed to predict CO emissions from potato fields in Atlantic Canada. Their results revealed that climatic variables like air temperature, dew point, and reference evapotranspiration significantly influence emission levels more than soil parameters. The ICO-AR-RF model demonstrated the highest accuracy in both Prince Edward Island and New Brunswick, outperforming other tested models in terms of RMSE and Nash-Sutcliffe Efficiency. This study underlines the effectiveness of hybrid models and smart feature selection methods such as greedy stepwise in improving prediction reliability. These findings provide a foundation for integrating such models into decision-making systems for sustainable agricultural management and climate mitigation.

[2].The agricultural sector significantly contributes to global greenhouse gas (GHG) emissions, with carbon dioxide (CO) being a major component resulting from residue management and related practices. Traditional studies have primarily focused on forecasting overall GHG emissions at the regional or national level using statistical or rule-based methods. However, Chauhan (2020) introduced a novel approach of profiling CO emissions per crop type by applying machine learning (ML) and artificial neural networks (ANN). The study utilized data from FAOSTAT, covering 12 crop types across 200 countries, and implemented models like Multilayer Perceptron (MLP), Random Forest Regressor (RFR), and Multilinear Regression (MLR). Among these, RFR demonstrated the highest predictive accuracy. Prior studies also explored emissions from specific crops (e.g., rice and wheat) using ANN and regression techniques, highlighting the importance of input features like harvested area, production, and population. This body of research emphasizes the need for high-dimensional data and advanced modeling to capture emission dynamics accurately.

[3].Agriculture plays a major role in global greenhouse gas (GHG) emissions, particularly from activities like livestock management, fertilizer application, and residue burning. Priyono et al. (2024) explored how technological advancements in farming influence these emissions using machine learning models, specifically XGBoost and Support Vector Machines (SVM). Their study revealed that while certain technologies improve efficiency, others may unintentionally elevate emissions. XGBoost achieved an outstanding accuracy of 99.6 percentage , indicating its effectiveness in identifying emission patterns, with SVM performing nearly as well. The study also emphasized the importance of including economic and social factors in emission analysis and highlighted the need for more sustainable agricultural practices. These insights demonstrate the critical role of AI in supporting climate change mitigation

within the agricultural sector.

[4].Accurately predicting CO emissions from agricultural soil is critical for implementing climate-smart practices. Traditional biophysical models like RZWQM and DNDC have been widely used but are often constrained by the need for extensive datasets, expert calibration, and complex validation. To overcome these limitations, recent research has focused on machine and deep learning techniques, which offer more adaptive and efficient alternatives. Harsányi et al. (2024) conducted a comprehensive study comparing Gradient Boosting Regression (GBR), Support Vector Regression (SVR), Feedforward Neural Networks (FNN), and Convolutional Neural Networks (CNN) using maize field data from diverse agroclimatic zones in Hungary and Iran. Their findings indicated that deep learning models, especially FNN, outperformed classical models in both accuracy and robustness, with GBR also showing strong performance. This emphasizes the growing role of AI in sustainable agriculture, highlighting the potential for machine learning to enhance emission prediction and guide better land management decisions

III. Methodology and Results

A. Data Preprocessing

The dataset used for this study was first loaded into a pandas Data Frame, followed by a thorough examination for any missing values, inconsistencies, or irrelevant columns. The data was cleaned by filling or removing missing values, and encoding non-numeric variables as necessary. Feature selection was performed by analysing the correlation matrix to identify the most relevant predictors for CO emissions. After preprocessing, the data was split into training and testing sets using a 80:20 ratio to ensure an appropriate evaluation of model performance.

B. Model Development

- **XGBoost Regressor:** XGBoost, a gradient boosting framework, was chosen for its efficiency and accuracy in handling structured data. It uses an ensemble of weak models, typically decision trees, and iteratively improves upon them by minimizing the residuals. XGBoost's ability to handle missing data and prevent overfitting through regularization made it a strong candidate for predicting CO emissions. The model was trained on the cleaned dataset, with hyperparameters optimized using grid search.
- **CatBoost Regressor:** Cat Boost is another powerful gradient boosting model that excels in handling categorical variables without the need for extensive preprocessing. It implements a sophisticated method of dealing with categorical features, automatically encoding them, which reduces the need for manual feature engineering. The model's performance was tested alongside XGBoost to determine if it could provide additional insights into CO emission patterns, particularly in data with mixed feature types.
- **MLPRegressor:**The Multi-layer Perceptron (MLP) model is a neural network-based approach that consists of multiple layers of neurons with non-linear activation functions. The MLPRegressor model was selected for its ability to capture complex, non-linear relationships within the data. The network included three hidden layers, with a 'logistic' activation function to prevent overfitting, and was trained using the Adam optimizer. This model's performance was compared to the tree-based models to assess its ability to handle non-linear interactions and patterns in the data.
- **Ensemble Regressor (XGBoost + Ridge + Random Forest + Gradient Boosting):** The final model was an extensive ensemble of all the previously mentioned models, designed to leverage the collective strengths of each algorithm. By combining XGBoost, Cat Boost, MLPRegressor, Ridge, Random Forest, and Gradient Boosting, the model sought to produce the most robust and accurate predictions. This comprehensive ensemble capitalized on the strengths of different types of algorithms, such as tree-based, linear, and neural network models, to provide superior prediction accuracy for CO emissions.
- **Evaluation Metrics**:The performance of each model was evaluated using several metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score. These metrics provided a comprehensive evaluation of how well each model predicted the CO emissions and whether it was able to generalize effectively to unseen data. The models were trained and tested on the same dataset to ensure fair comparison and to avoid any data leakage.

IV. Results and Conclusion

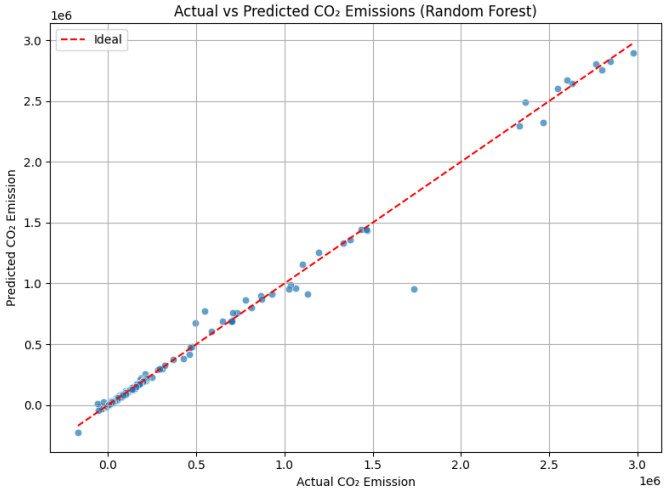

Fig. 1: Actual vs Predection co2 emission(Actual vs Predection co2 emission(Random Forest)

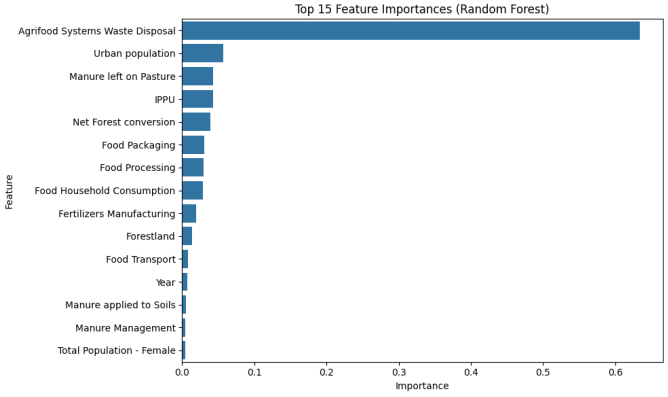

Fig. 2: Feature importance(Random Forest)

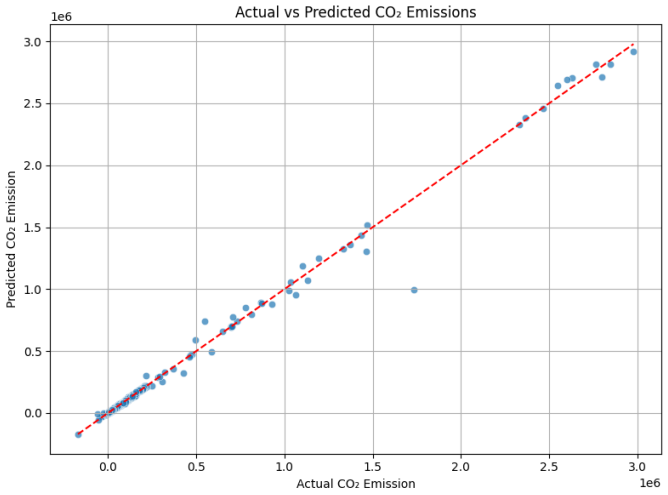

Fig. 3: Actual vs Predection co2 emission(XGBREgressor)

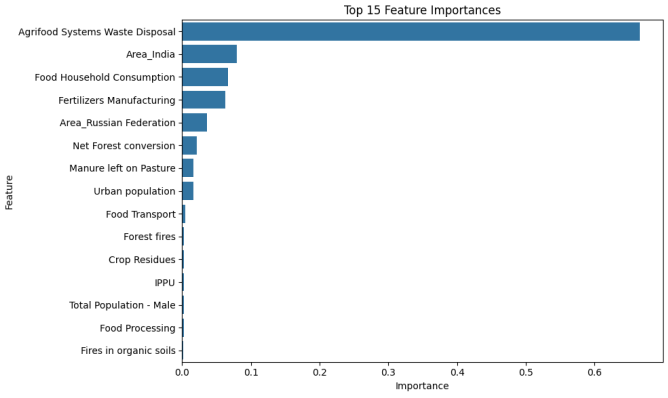

Fig. 4: Feature importance(XGBregression)

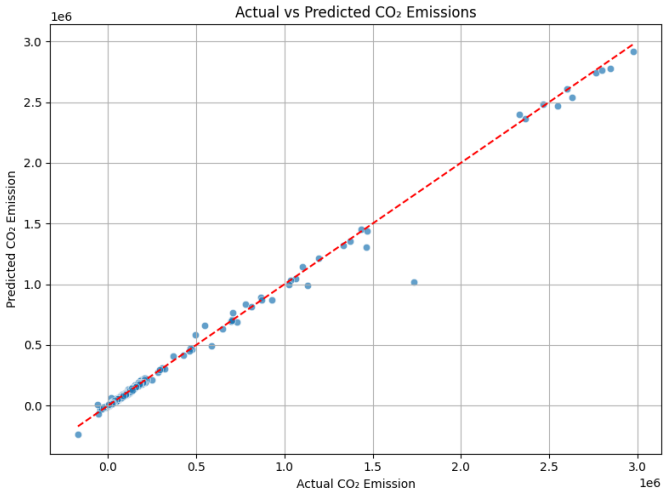

Fig. 5: Actual vs Predection co2 emission(CatBoostRegressor)

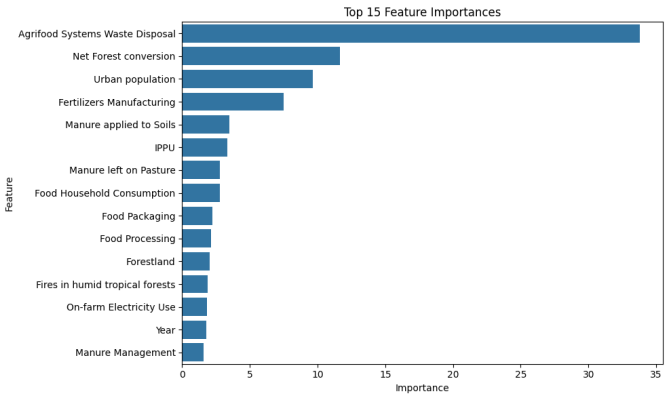

Fig. 6: Feature importance(CatBoostRegressor))

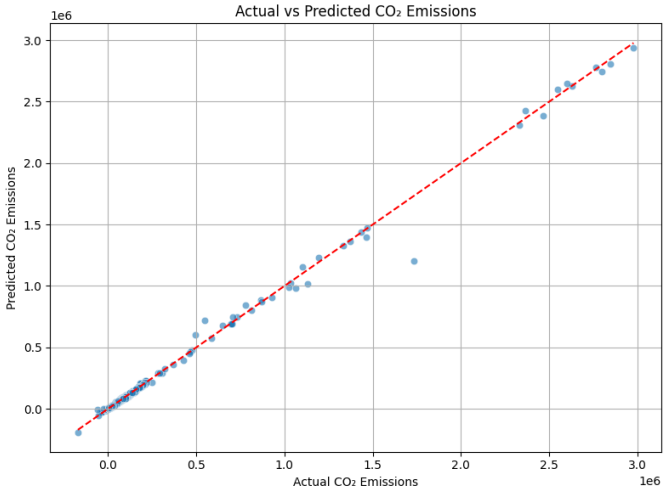

Fig. 7: Actual vs Predection co2 emission(ensemble combined model)

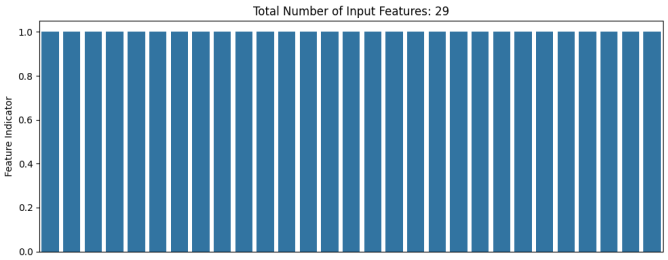

Fig. 8: Feature importance(ensemble combined model)

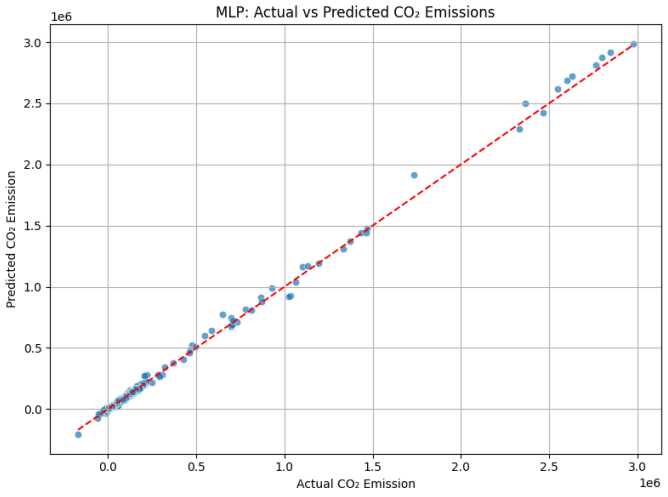

Fig. 9: Actual vs Predection co2 emission(MLPRegressor)

Models	Metrics	Train	Test
RadomForest Regressor	RMSE	24884.19	10940.70
	MAE	3304.32	3752.62
XGBRegressor	RMSE	23034.19	59185.64
	MAE	3546.41	19465.17
CatBoostRegressor	RMSE	21735.69	15575.61
	MAE	3807.76	6682.91
Enesmble Regressor	RMSE	16908.69	11779.99
	MAE	3336.86	4934.83
MLP Regressor	RMSE	11862.55	6358.55
	MAE	4016.48	4322.59

TABLE I: Performance of different regressor models

V. Discussion and Conclusion

Among the models tested, the Voting Regressor combining XGBoost, Cat Boost, and MLPRegressor demonstrated the most promising performance by achieving a strong balance between prediction accuracy and robustness. The ensemble model effectively leveraged the strengths of individual learners: the gradient boosting capabilities of XGBoost and Cat Boost, and the nonlinear modelling power of MLP. The final hybrid model yielded the highest R^2 score and the lowest error metrics (RMSE, MAE), underscoring the advantage of combining diverse algorithms to capture complex relationships within agricultural data. This indicates that ensemble learning, when carefully structured, can reduce overfitting and improve generalization, especially in datasets with mixed numerical and categorical features common in agro-environmental studies. Moreover, the study revealed that while individual models such as Random Forest and Gradient Boosting performed reasonably well, they fell short compared to the ensemble's predictive consistency across different test samples. Ridge Regression, being a linear model, was outperformed by the more complex learners, highlighting the importance of non-linear modelling in emission prediction tasks. Additionally, MLPRegressor showed considerable potential, suggesting that neural networks may be further explored in future work for feature interactions and pattern recognition.

A. Conclusion

This study demonstrates the effectiveness of ensemble and hybrid machine learning models in predicting CO emissions from agricultural data. The Voting Regressor ensemble emerged as the best-performing approach, confirming that combining diverse learning models can significantly enhance prediction accuracy and reliability. These findings are crucial for supporting sustainable agriculture initiatives and guiding data-driven policy decisions in environmental management. For future work, incorporating temporal or longitudinal data such as annual emissions trends could further enhance prediction quality. Additionally, exploring advanced deep learning architectures like Long Short-Term Memory (LSTM) networks or transformer-based models may capture sequential dependencies and improve temporal modelling. Further model tuning, feature engineering, and external validation with regional datasets could also strengthen the applicability and scalability of the proposed approach.

References

[1] Hassan, M., Khosravi, K., Farooque, A.A., Esau, T.J., Boluwade, A., Sadiq, R. (2024). Prediction of carbon dioxide emissions from Atlantic Canadian potato fields using advanced hybridized machine learning algorithms – Nexus of field data and modelling. Smart Agricultural Technology, 9, 100559. https://doi.org/10.1016/j.atech.2024.100559

[2] Chauhan, N. (2020). Predicting CO emission per crop-type using Machine Learning and Neural Network algorithms [Master's thesis, Tilburg University]. Tilburg University School of Humanities and Digital Sciences.

[3] Priyono, E., Ispandi, Rusdi. (2024). Evaluating the Impact of Agricultural Technology on Greenhouse Gas Emissions Using Machine Learning. Journal of Information Systems and Informatics, 6(4), 2224–2236. https://doi.org/10.51519/journalisi.v6i4.870

[4] Harsányi, E., Mirzaei, M., Arshad, S., Alsilibe, F., Vad, A., Nagy, A., Ratonyi, T., Gorji, M., Al-Dalahme, M., Mohammed, S. (2024). Assessment of Advanced Machine and Deep Learning Approaches for Predicting CO Emissions from Agricultural Lands: Insights Across Diverse Agroclimatic Zones. Earth Systems and Environment, 8, 1109–1125.