INFO-I 513 - Usable Artificial Intelligence

Surya Teja Mothukuri – smothuk@iu.edu

Final Project Report – Carbon Footprint Detection

Introduction :

Carbon Emissions are a key contributor to climate change, mainly through the production of greenhouse gases like CO_2, it is important to comprehend them. We can learn more about which routine activities, including food, travel, and energy consumption, have the most environmental effect by monitoring and analyzing carbon emissions at the person or household level. People, groups, and legislators are better equipped to make decisions that promote sustainability and lessen global warming because of this information.

Objectives :

- To perform intensive data exploration and data visualization methods to the newly
- engineered dataset.
- Performed statistical analysis using ANOVA and Chi-Square tests to identify significant differences in emissions
- To preprocess the dataset using StandardScaler and label encoding, ensuring all features are transformed appropriately to fit the models without data leakage.
- To implement and evaluate multiple machine learning algorithms, including Ridge Regression, Decision Tree, Support Vector Regression, Random Forest, and XGBoost, using cross-validation and error metrics like RMSE, MAE, and R^2.
- To design and train a deep learning model (ANN) using TensorFlow/Keras to capture complex nonlinear relationships in the data and compare its performance with traditional ML models.
- To develop an interactive Streamlit dashboard, allowing users to input personal lifestyle details and receive real-time carbon emission predictions using the trained XGBoost model and saved preprocessing pipeline.

Motivation :

I am a firm believer in leaving this world in a better place than we found it. Carbon emissions are linked directly to global warming and its consequences are the worst – rising sea levels, warm temperatures, bio-diversity loss etc. Tracking our own emissions becomes of utmost importance in these tough times and adverse effects. This helps us identify actionable changes in our lifestyle, what to reduce and what to not have in our lifestyle. Insights from this analysis will help people question their actions in terms of 'Is this behavior of mine really that necessary?' and can guide personal decisions. Climate change, food and water insecurity will affect everyone, regardless of geography and lifestyle. By understanding this analysis, society and policymakers can make informed and calculated decisions and encourage responsible consumption. This will help in building a more resilient and environmentally responsible future.

Materials and Method :

Dataset :

Here, I will be using the open-source Carbon Emissions Dataset which consists of structured data and information about lifestyle habits related to daily consumption and activities that contribute to carbon emissions such as information about demographics, lifestyle, energy consumptions, diet habits, water usage, waste management etc.

Work Done/ Preliminary Report :

In this project, I cleaned the dataset by handling missing values and removing redundancies, followed by engineering new emission-related variables—namely diet_emissions, vehicle_emissions, waste_emissions, and flight_emissions—using real-world CO_2 equivalence mappings. I conducted statistical analyses such as ANOVA and Chi-Square tests to uncover significant differences and associations among lifestyle-based categorical variables. A range of visualizations was applied to the engineered dataset to explore patterns in carbon emissions, including boxplots comparing emissions across diet, transport, heating sources, and air travel frequency; a correlation heatmap for numerical emission features; a mosaic plot examining intersections between diet, transport, and recycling levels; radar charts comparing emission profiles across emission tiers, gender, and screen time levels; and stacked bar charts highlighting the average breakdown of emission types by transport mode. These methods helped reveal the key drivers of individual carbon footprints across various behavioral dimensions.

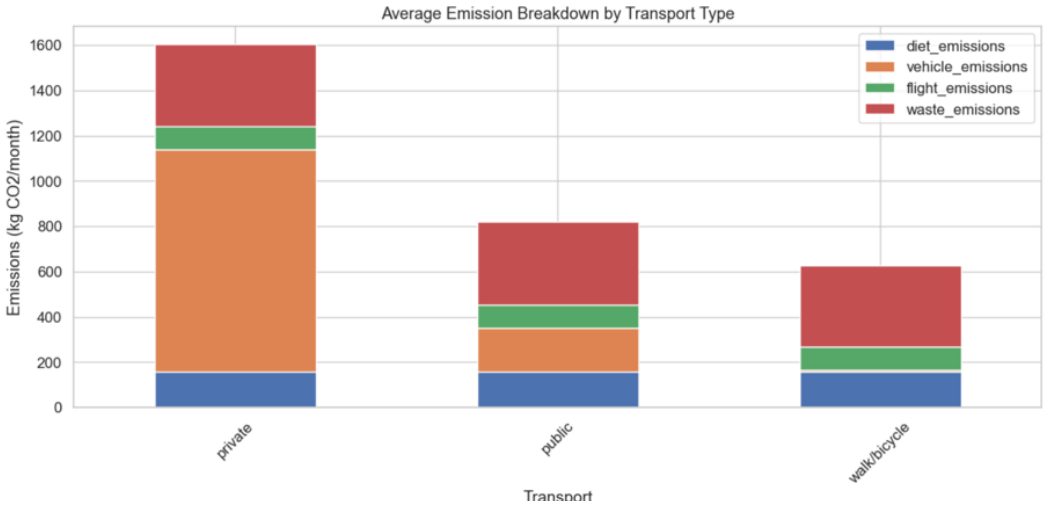
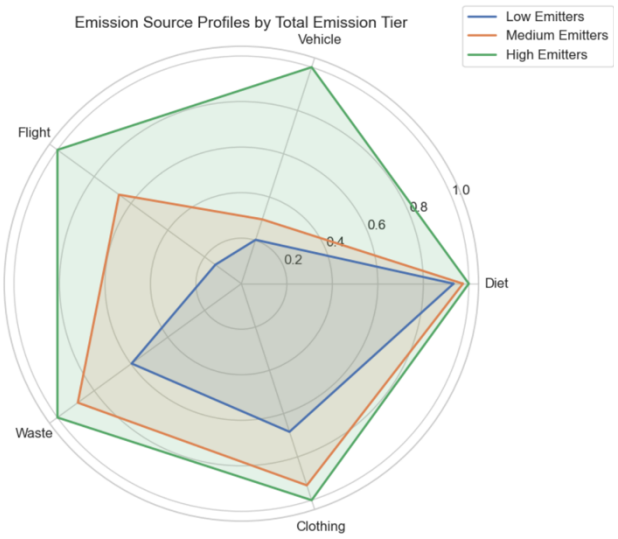


Fig 1. Avg Emission Breakdown by Transport Type

This chart shows that individuals using private transport have the highest carbon emissions, mainly due to vehicle-related output. In contrast, those who walk, bike, or use public transport have much lower total emissions. While waste and diet emissions remain consistent across groups, vehicle use is the major factor driving differences. This highlights how shifting to sustainable transport can greatly reduce overall carbon footprint.



This spider plot groups individuals into Low, Medium, and High Emitters based on total carbon emissions and compares their average emissions across five sources. High Emitters show the highest values across all categories, especially in vehicle and flight emissions. Medium Emitters are more balanced, while Low Emitters have significantly lower emissions in all areas except diet, which remains consistent. This suggests that transportation and consumption habits are key drivers of high carbon footprints.

Fig 2. Emission Profiles by Total Emission Tiers

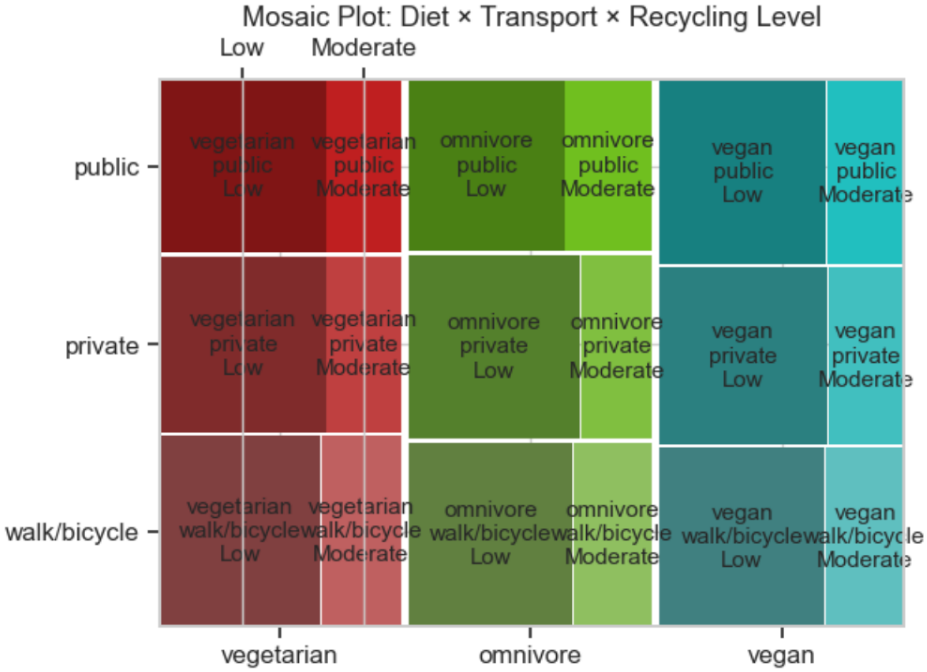
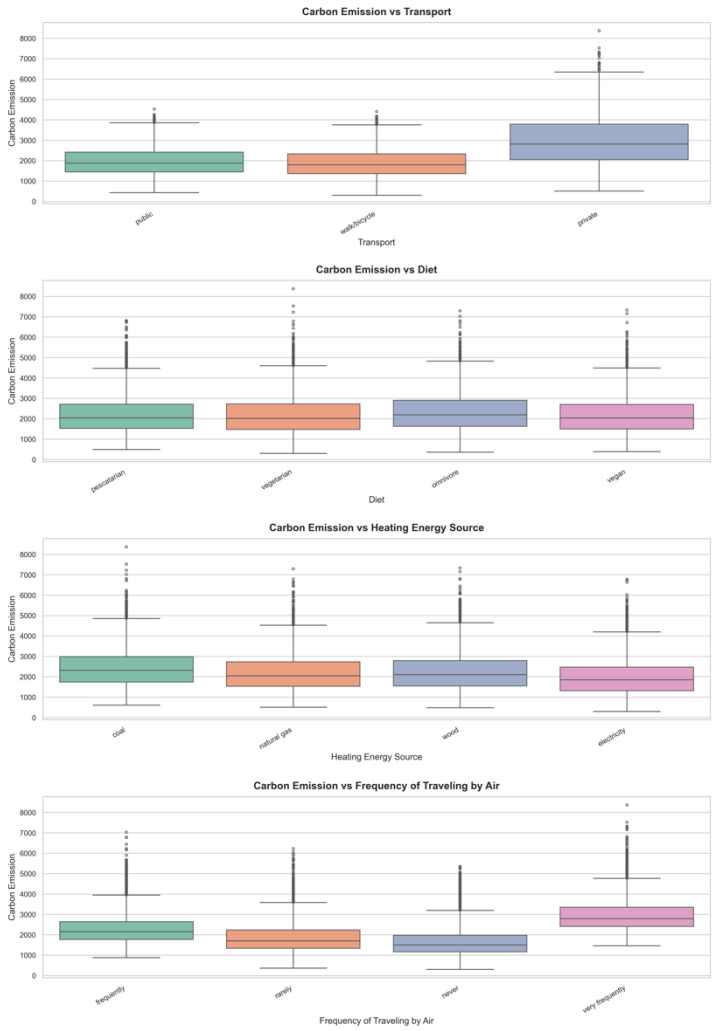


Fig 3. Mosaic Plot of Diet x Transport x Recycling Level

This mosaic plot shows how diet, transport, and recycling levels intersect. Vegan users who use public or walk/bicycle transport with low to moderate recycling are the most, suggesting this is the most common eco-conscious group. Omnivores with private transport tend to recycle less, reflecting a less sustainable pattern. Overall, the plot highlights clear lifestyle clusters based on environmental behavior.



Transport :

- Individuals using private transport show notably higher carbon emissions with more outliers and a higher median, compared to those using public or walking/bicycling.

Diet :

- People with a meat-heavy diet tend to have the highest emissions, while vegetarian and vegan diets show lower median values, reinforcing the environmental impact of dietary choices.

Heating Energy Source:

- Those using electric or natural gas as heating sources tend to have higher emissions than those using wood or renewables, with coal surprisingly lower, likely due to sample size or mixed heating methods.

Travelling by air :

- Frequent and very frequent flyers exhibit the highest carbon emissions, while those who rarely or never fly have lower medians and fewer extreme values.

Fig 4. Boxplots of certain categorical columns

We categorized individuals into Low, Medium, and High emitters using pd.qcut and analyzed the average emissions from various sources. The plot shows that vehicle emissions are the largest contributor for high emitters, followed by waste and clothing emissions, which also increase across groups. Flight emissions rise slightly, while diet emissions remain consistent, suggesting that transport and consumption habits are key drivers of high carbon output.

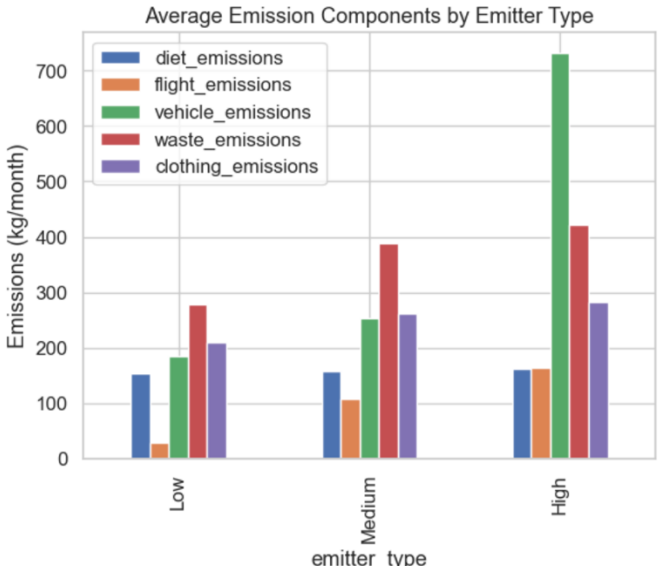
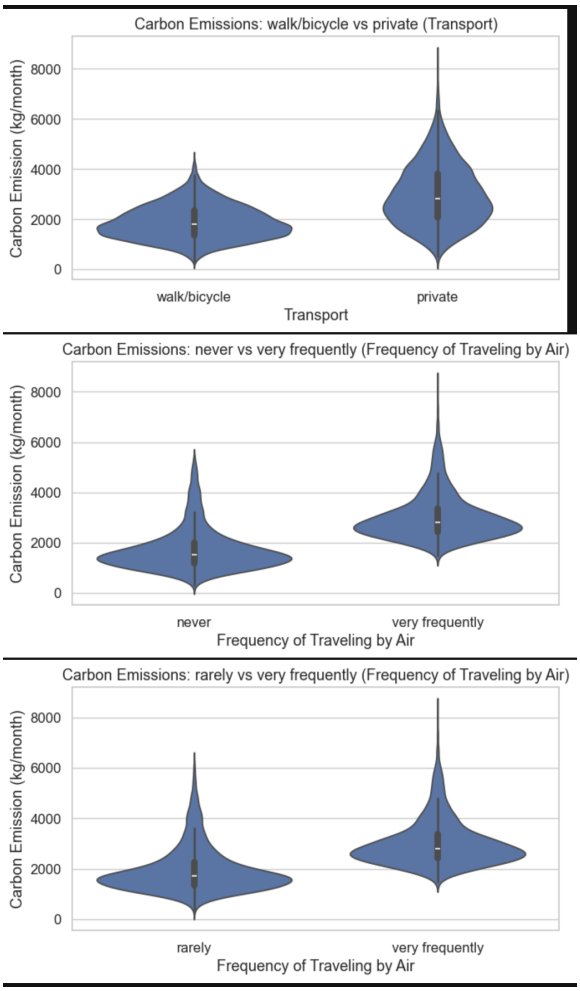


Fig 5. Avg. Emission Components by Emitter Type

Taking feedback from Prof. Silva, I have implemented the Mann–Whitney U test to statistically compare carbon emission distributions between different groups within categorical variables.



We applied the **Mann–Whitney U test**, a non-parametric statistical test that compares the distributions of two independent samples. The test was conducted on all valid pairwise combinations of categorical variable groups. Several variables exhibited statistically significant differences ($p < 0.05$), including **transport mode**, **frequency of traveling by air**, **body type**, **heating energy source**, and **sex**.

For instance, individuals using **private transport** showed significantly higher emissions than those who walked or biked, and people who **traveled by air very frequently** emitted much more than those who rarely or never flew. The extremely low p-values (e.g., 1e-287 for public vs. private transport) strongly support these differences. These findings were further illustrated through **violin plots**, which visually confirmed the higher and more skewed emission distributions among high-impact groups, reinforcing the statistical results.

Fig 6. Carbon Emissions Across Statistically Significant Group Pairs

Variable	Group A	Group B	U-Statistic	P-Value	Significant
Transport	walk/bicycle	private	2477968.0	0.0	True
Frequency of Traveling by Air	never	very frequently	699883.0	0.0	True
Frequency of Traveling by Air	rarely	very frequently	980807.5	0.0	True
Transport	public	private	2615512.0	5.078568e-287	True
Frequency of Traveling by Air	frequently	never	4696622.5	3.464515e-216	True
Frequency of Traveling by Air	frequently	very frequently	1615840.5	4.304828e-205	True
Frequency of Traveling by Air	frequently	rarely	4243425.5	3.256948e-106	True
Body Type	obese	underweight	4130679.5	1.929315e-76	True
Heating Energy Source	coal	electricity	4135657.5	1.120981e-71	True
Sex	female	male	10155931.0	2.662929e-59	True

Table 1. Top Significant Group Comparisons from Mann–Whitney U Test on Carbon Emissions

To identify the most effective regression model for predicting carbon emissions, we implemented a grid search-based hyperparameter tuning process using GridSearchCV from Scikit-learn. Six models were evaluated: **Ridge Regression, Decision Tree Regressor, Support Vector Regressor (SVR), Random Forest Regressor, XGBoost Regressor,** and **CatBoost Regressor**. For each model, a predefined grid of hyperparameters was constructed based on commonly tuned parameters such as max_depth, learning_rate, n_estimators, and others, specific to each algorithm's requirements.

Each model was trained and validated using 3-fold cross-validation, and performance was measured using the Root Mean Squared Error (RMSE). GridSearchCV automatically evaluated all hyperparameter combinations and selected the best configuration for each model. The best-performing version of each model was stored for later evaluation. This process ensured that all models were optimized fairly under the same validation strategy, allowing a robust comparison of their predictive performance on the carbon emission dataset.

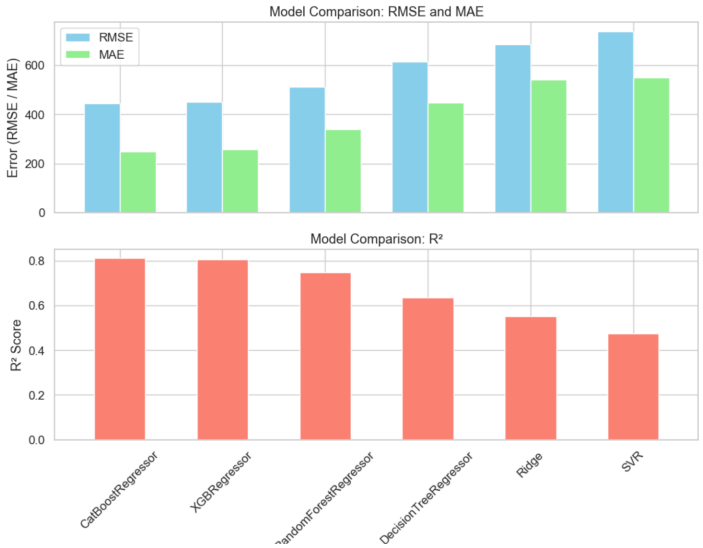
```
Running GridSearchCV for Ridge...
Best Parameters: {'alpha': 10.0}
Best RMSE: 692.7813

Running GridSearchCV for DecisionTreeRegressor...
Best Parameters: {'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 10}
Best RMSE: 627.8932

Running GridSearchCV for SVR...
Best Parameters: {'C': 10, 'epsilon': 0.1, 'gamma': 'auto', 'kernel': 'rbf'}
Best RMSE: 787.8949

Running GridSearchCV for RandomForestRegressor...
Best Parameters: {'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
Best RMSE: 533.6153

Running GridSearchCV for XGBRegressor...
Best Parameters: {'colsample_bytree': 0.8, 'learning_rate': 0.3, 'max_depth': 3, 'n_estimators': 50, 'subsample': 1.0}
Best RMSE: 465.5053

Running GridSearchCV for CatBoostRegressor...
0:      learn: 987.5703406      total: 57.5ms   remaining: 11.5s
100:    learn: 446.3019181      total: 133ms    remaining: 130ms
199:    learn: 387.2428546      total: 203ms    remaining: 0us
Best Parameters: {'depth': 6, 'iterations': 200, 'learning_rate': 0.05}
Best RMSE: 460.0998
```

Fig 7. GridSearchCV Results

Model	RMSE	MAE	R-squared
Ridge	683.631	541.0233	0.5505
Decision Tree	616.1371	448.1329	0.6349
SVR	739.1648	551.227	0.4745
Random Forest	511.2302	338.6591	0.7486
XGBRegressor	450.1356	256.0065	0.8051
CatBoostRegressor	443.5909	247.3343	0.8107

Table 2. Regression Model Results and its metrics

The model comparison table shows that CatBoost Regressor and XGBoost Regressor significantly outperform the other models across all three evaluation metrics: RMSE, MAE, and R^2. CatBoost achieved the lowest RMSE (443.59) and lowest MAE (247.33), indicating the most accurate predictions with the smallest average error. It also achieved the highest R^2 score (0.8107), suggesting it explains over 81% of the variance in carbon emissions.



XGBoost closely followed, with slightly higher error but still strong generalization performance ($R^2 = 0.8051$). Ensemble tree-based models clearly outperform linear (Ridge) and non-linear (SVR) models, highlighting the presence of complex, non-linear relationships in the dataset. The relatively poor performance of SVR (highest RMSE and lowest $R^2 = 0.4745$) further confirms that simpler models may not adequately capture the patterns in this feature-rich dataset.

Fig 8. Metrics Comparisons for all 6 models

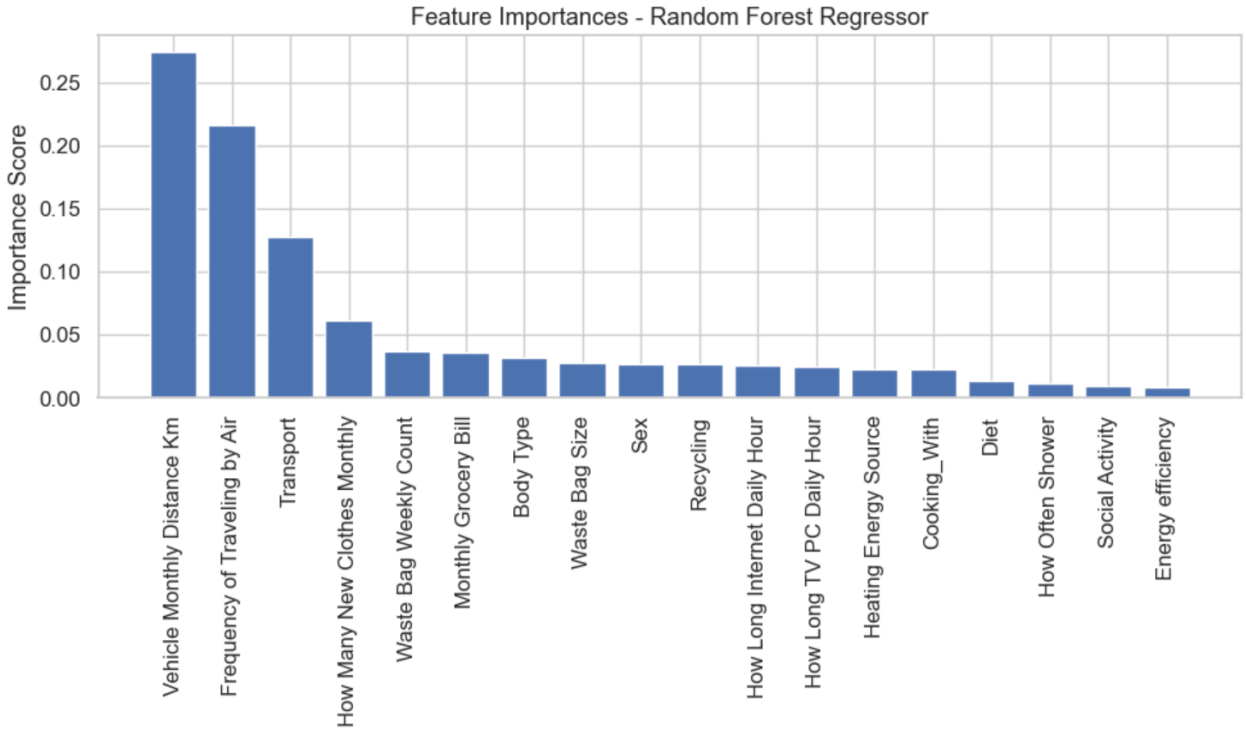


Fig 9. Feature Importances for Random Forest Regressor

The feature importance analysis from the Random Forest Regressor reveals that “Vehicle Monthly Distance Km” is the most influential predictor of carbon emissions, contributing nearly 28% to the overall model decision-making. This is followed by “Frequency of Traveling by Air” and “Transport”, which also show strong influence, highlighting the significant role of transportation habits in determining individual carbon footprints. Other moderately important features include “How Many New Clothes Monthly”, “Waste Bag Weekly Count”, and “Monthly Grocery Bill”, suggesting that consumer behavior and waste generation also contribute meaningfully to emissions.

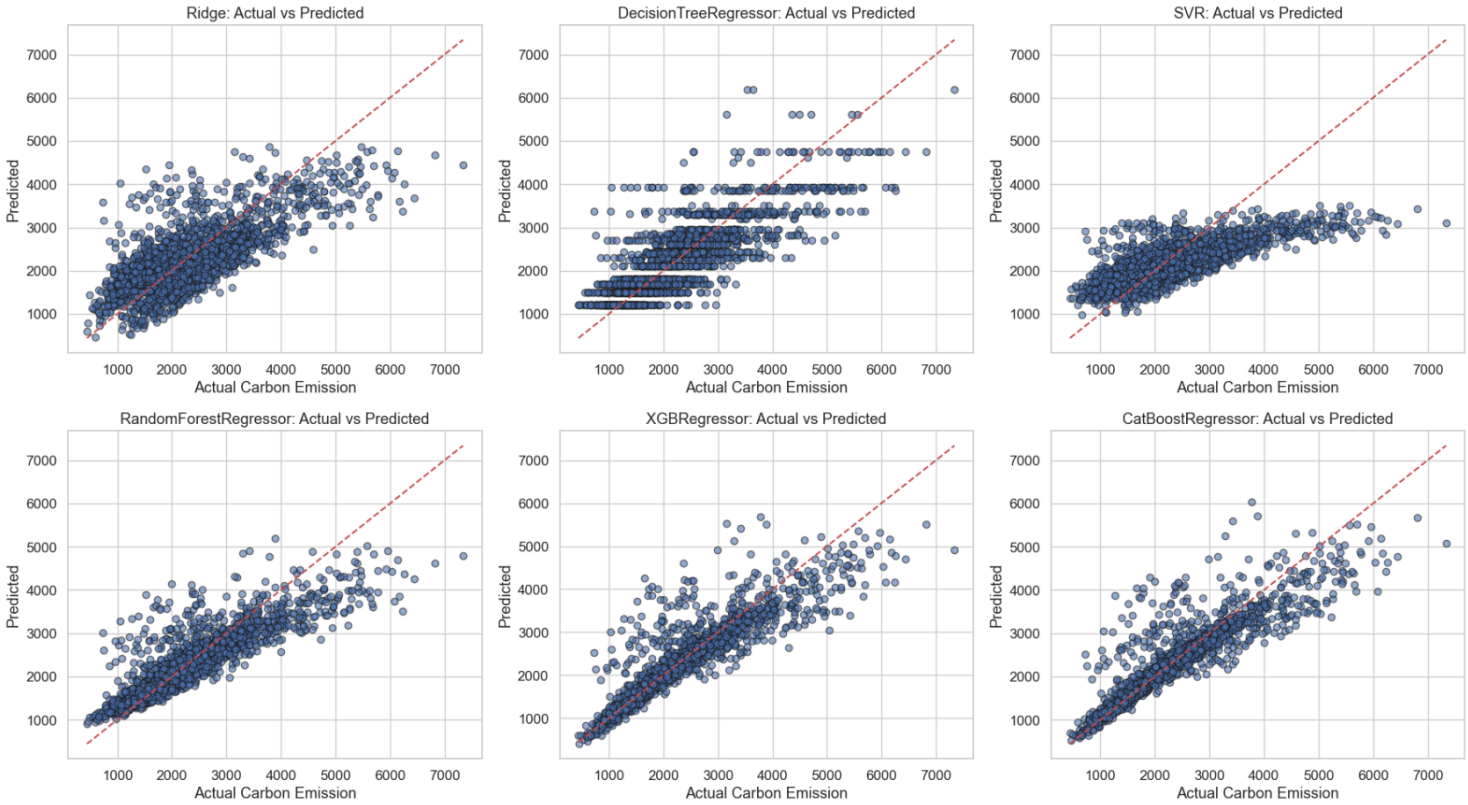


Fig 10. Actual vs Predicted for all 6 models

The actual vs. predicted scatter plots provide a visual comparison of model performance, where ideal predictions lie along the diagonal red line. Among the models, CatBoost Regressor and XGBoost Regressor show the tightest clustering around the diagonal, indicating high accuracy and minimal deviation in predictions. This visually supports their strong performance metrics (lowest RMSE and highest R^2).

The Random Forest Regressor also performs well, though with slightly more spread. Ridge Regression shows a linear trend but with wider dispersion, while SVR tends to underpredict at higher emission values. Decision Tree Regressor exhibits clear overfitting behavior, seen in its stair-step predictions and reduced generalization. These plots clearly demonstrate that ensemble-based gradient boosting models outperform simpler linear and tree-based models in capturing the complexity of carbon emission patterns.

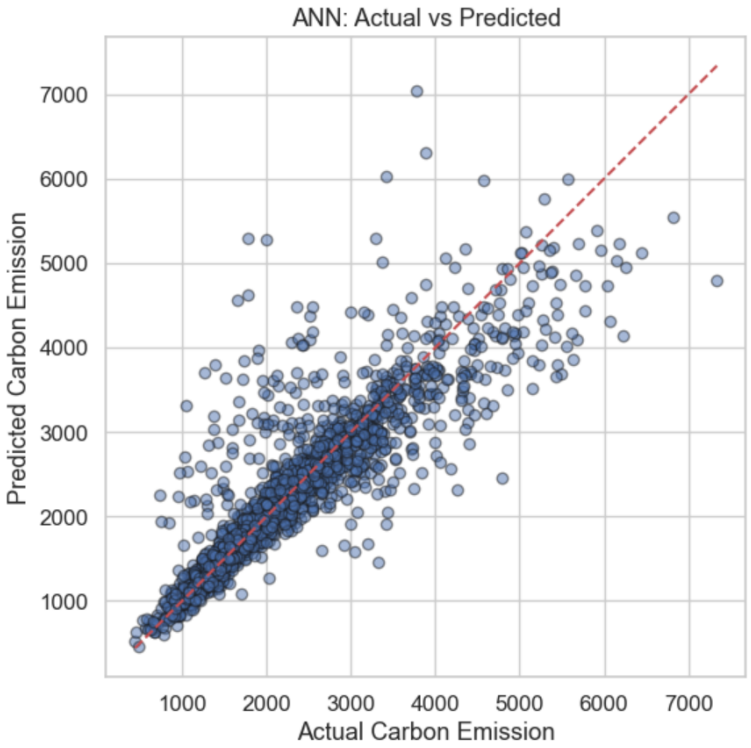


Fig 11. Actual vs Predicted for ANN Model

The scatter plot of **ANN predictions vs. actual carbon emissions** shows a strong alignment along the diagonal reference line, indicating that the neural network effectively captured the underlying patterns in the data. The high density of points closely following the diagonal suggests that most predictions are accurate, particularly in the mid-range of emission values.

While some deviation and underestimation occur at higher emission values (above 5000 kg/month), the overall trend remains consistent, reflecting the model's ability to generalize well. Compared to traditional ML models, the ANN demonstrates competitive performance, benefiting from its capacity to model complex, non-linear relationships. This visual evidence reinforces the ANN's validity as a robust alternative to ensemble tree-based methods in this task.

StreamLit Dashboard (code separately in app.py) :

To make our carbon emission prediction model accessible and user-friendly, we developed an **interactive dashboard using Streamlit**, a lightweight Python library for building data apps. After evaluating multiple machine learning models, **CatBoost Regressor** was identified as the best-performing algorithm based on its lowest RMSE (443.59), lowest MAE (247.33), and highest R^2 score (0.8107). We saved the trained CatBoost model using joblib and loaded it into the Streamlit app to generate real-time predictions.

The dashboard allows users to input 18 lifestyle and demographic features such as body type, transport mode, air travel frequency, waste habits, and internet usage. Each input field is implemented through intuitive UI elements like dropdowns and sliders. Upon submitting the inputs, the app scales the data using the same StandardScaler used during training and feeds it into the CatBoost model. The predicted **monthly carbon emission** (in kg CO_2) is displayed instantly, making the tool both informative and engaging for users interested in understanding their environmental impact.

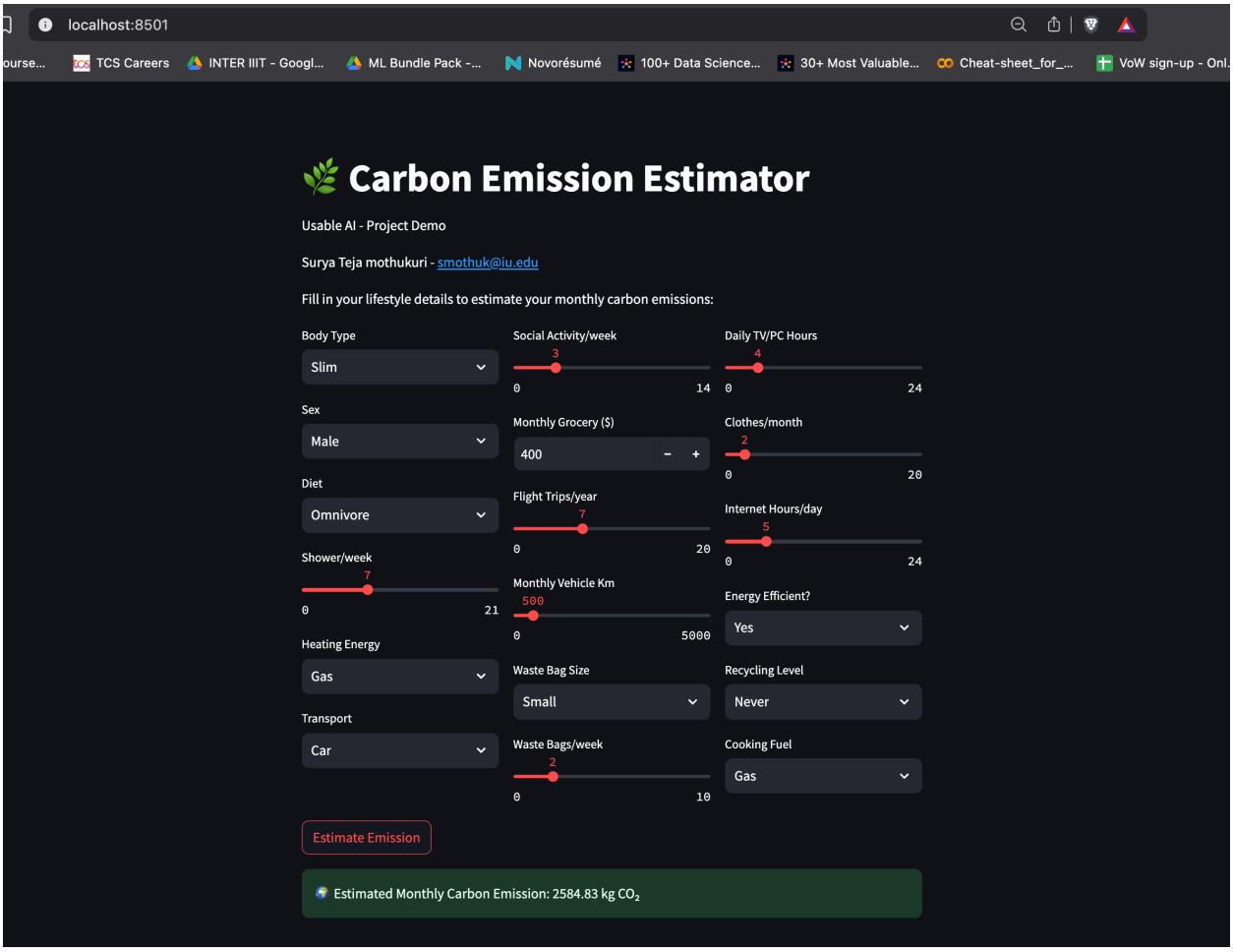


Fig 12. Streamlit Dashboard

Discussion/Conclusion :

This project successfully explored the use of machine learning and deep learning to predict individual carbon emissions based on lifestyle behaviors. Through extensive preprocessing, feature engineering, and visualization, we identified key contributors to emissions—most notably transportation habits such as monthly vehicle usage and air travel frequency. Statistical tests, like the Mann–Whitney U test, confirmed significant emission differences across categorical groups such as transport mode, heating energy source, and dietary habits. These insights emphasized that sustainable choices in travel and consumption can meaningfully reduce personal carbon footprints.

Among the six machine learning models evaluated, CatBoost Regressor emerged as the most effective, achieving the lowest RMSE and MAE, and the highest R^2 (0.8107), closely followed by XGBoost. Deep learning via ANN also produced competitive results, demonstrating its capacity to model non-linear patterns in emissions data. The project culminated in the development of an interactive Streamlit dashboard, allowing users to input their lifestyle data and receive real-time emission estimates using the trained CatBoost model. This application not only empowers individuals to better understand their environmental impact but also supports informed decision-making for more sustainable living.