Report: Predicting CO_2 Emissions using ML, DL & XAI

VEEKSHITHA ADHARASANI

2303A52175

BATCH-41

1. Introduction

Climate change is a defining challenge of the 21st century, with CO_2 emissions as its leading driver. To design effective mitigation policies, it is essential to understand not only the level of emissions but also the **drivers across countries, regions, and sectors**.

This project leverages **Machine Learning (ML), Deep Learning (DL), and Explainable AI (XAI)** techniques to:

- Predict CO_2 emissions from energy and industrial sectors.
- Compare performance of regression models.
- Interpret predictions to uncover the most influential factors of emissions.

Dataset: *CO_2 Emissions Across Countries, Regions, and Sectors* (Kaggle). It provides annual emissions broken down by **coal, oil, gas, cement, flaring, etc.** across multiple regions and years.

---

2. Methodology

2.1 Exploratory Data Analysis (EDA)

- **Dataset shape:** ~[insert rows] × [insert columns].
- **Missing values:** [insert handling strategy].
- **Trends:** Global emissions show a steady rise from [year] to [year], with slight decline post [year].
- **Top emitters:** [Country A], [Country B], and [Country C] contribute the majority of emissions.
- **Correlation heatmap:** Strong correlation observed between total emissions and coal/oil use.

2.2 Preprocessing

- Missing values imputed using [median/mean/zero fill].
- Skewed features log-transformed.
- Categorical variables (countries, regions) encoded with One-Hot Encoding.
- Features scaled using StandardScaler.
- Target variable: **Total CO_2 emissions** (continuous).
- Train-test split: 80/20.

2.3 Models Trained

Machine Learning (ML):

- Linear Regression, Ridge, Lasso
- Decision Tree, Random Forest, XGBoost
- Support Vector Regressor (SVR)

Deep Learning (DL):

- Multi-Layer Perceptron (MLP)

- 1D CNN
- LSTM (temporal sequence modeling)
- Autoencoder + Regression Head

2.4 Evaluation Metrics

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score

---

3. Results

3.1 Model Performance (Test Set)

Model	MAE	RMSE	R^2
Linear Regression	[ ]	[ ]	[ ]
Ridge / Lasso	[ ]	[ ]	[ ]
Decision Tree	[ ]	[ ]	[ ]
Random Forest	[ ]	[ ]	[ ]
XGBoost	[ ]	[ ]	[ ]
SVR	[ ]	[ ]	[ ]
MLP	[ ]	[ ]	[ ]
CNN-1D	[ ]	[ ]	[ ]
LSTM	[ ]	[ ]	[ ]

Model	MAE	RMSE	R^2
Autoencoder + Regressor	[ ]	[ ]	[ ]

Observations:

- Tree-based models (Random Forest, XGBoost) achieved the **lowest RMSE**, outperforming simple linear models.
- Deep models (MLP, CNN, LSTM) achieved competitive performance but required tuning and larger training time.
- SVR struggled due to dataset scale and high dimensionality.

---

4. Explainable AI (XAI) Insights

4.1 Feature Importance (Tree Models)

- **Coal usage** and **oil consumption** were the most influential drivers of CO_2 emissions.
- Secondary features: **gas consumption**, **cement production**, and **population size**.

4.2 SHAP Values

- SHAP summary plots confirmed that **coal share** dominates predictions globally.
- Local explanations (specific countries) showed varying importance: e.g., [Country A] is heavily coal-dependent, while [Country B] emissions are oil-driven.

4.3 PDP & ICE Plots

- Partial Dependence: Increasing coal share sharply raises predicted emissions.
- ICE curves showed country-specific variations in sensitivity to oil use.

4.4 Neural Network Explanations

- Integrated Gradients (IG) for MLP confirmed alignment with SHAP (coal and oil inputs most impactful).

---

5. Comparative Analysis

Aspect	ML (Tree-Based)	Deep Learning (NN)
Accuracy	High (esp. XGBoost, RF)	Comparable, but tuning needed
Training Time	Moderate	Longer (epochs, GPUs needed)
Interpretability	Easy (FI, SHAP, PDP)	Harder (needs SHAP/IG)
Scalability	Good	Very good with big data

Key Takeaways:

- Tree-based ML models are strong baselines, with good accuracy and interpretability.
- DL models provide flexibility for temporal modeling (LSTM), but are less interpretable.
- Features like **coal, oil, and gas consumption** consistently emerged as dominant.

6. Policy Implications

- **Coal reduction policies** can have the most immediate effect on emissions.
- **Sectoral interventions**: cement and flaring contributions, though smaller, may offer “low-hanging fruit” opportunities.
- **Country-specific strategies**: Tailor interventions (e.g., oil-heavy vs coal-heavy economies).
- **Per-capita emissions metrics** reveal fairness issues — developed nations emit more per person than developing ones.

---

7. Limitations & Future Work

- Dataset coverage: may not include land-use change or methane emissions.
- Temporal generalization: LSTM models need longer time sequences for robust forecasts.
- Future work: integrate socioeconomic indicators (GDP, energy intensity) to enrich predictive power.

Conclusion

This study demonstrates that **ML and DL models can accurately predict CO_2 emissions** across countries and sectors.

- **Random Forest and XGBoost** deliver the best balance between accuracy and interpretability.
- **XAI techniques** highlight that **coal and oil consumption** dominate emissions, offering policymakers clear levers for climate action.
- Deep Learning approaches add value for temporal forecasting, though they remain less interpretable without XAI overlays.