Assignment 5

Syeda hafsa fathima

2303A52381

BATCH 39

Predicting CO_2 Emissions using ML, DL & XAI

1. **Introduction**

- The 21st century is characterized by climate change, the primary force of which is the CO 2 emissions. In order to develop effective mitigation policies, the level of emissions is not sufficient; the drivers within the countries, regions and sectors should be known.
- The given project uses the approaches of Machine Learning (ML), Deep Learning (DL), and Explainable AI (XAI) to:
- Forecast energy and industrial CO 2 emissions.
- Compare the performance of regression models.
- Make predictions to be able to find out the most impactful elements of emissions.
- Dataset: CO 2 Emissions by Countries, Region, and Sectors (Kaggle). It gives annual emissions by coal, oil, gas, cement, flaring etc. by various 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	11.06	17.34	0.91
Random Forest	1.79	2.79	0.99
XGBoost	1.84	2.90	0.99
Support Vector Regressor	7.10	15.62	0.92
Deep MLP	7.64	11.17	0.96
CNN	3.28	4.60	0.99
LSTM	2.60	4.20	0.91
CNN + LSTM (Hybrid)	2.50	4.10	0.92
Autoencoder + Regression	2.85	4.55	0.88

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.