Predicting CO_2 Emissions for Climate Action Using Machine Learning

Introduction

Climate change, a pressing global challenge, is addressed under the United Nations' Sustainable Development Goal 13 (SDG 13): Climate Action. Rising CO_2 emissions from human activities threaten ecosystems and livelihoods, necessitating data-driven solutions. This project leverages machine learning to predict CO_2 emissions, using features like population, GDP, primary energy consumption, and industry-related emissions, to inform sustainable policies.

Methodology

The project utilized a dataset [insert source, e.g., synthetic data generated for 2010–2020 or Our World in Data] with variables: population, gdp, primary_energy_consumption, oil_co2, cement_co2, and other_industry_co2, targeting co2 as the dependent variable. Two supervised learning models—Linear Regression and Random Forest—were trained, with Random Forest outperforming (MAE: 332.04, R^2: 0.8142) compared to Linear Regression (MAE: 465.39, R^2: 0.7597). Feature importance analysis from Random Forest revealed key drivers, while visualizations (e.g.,

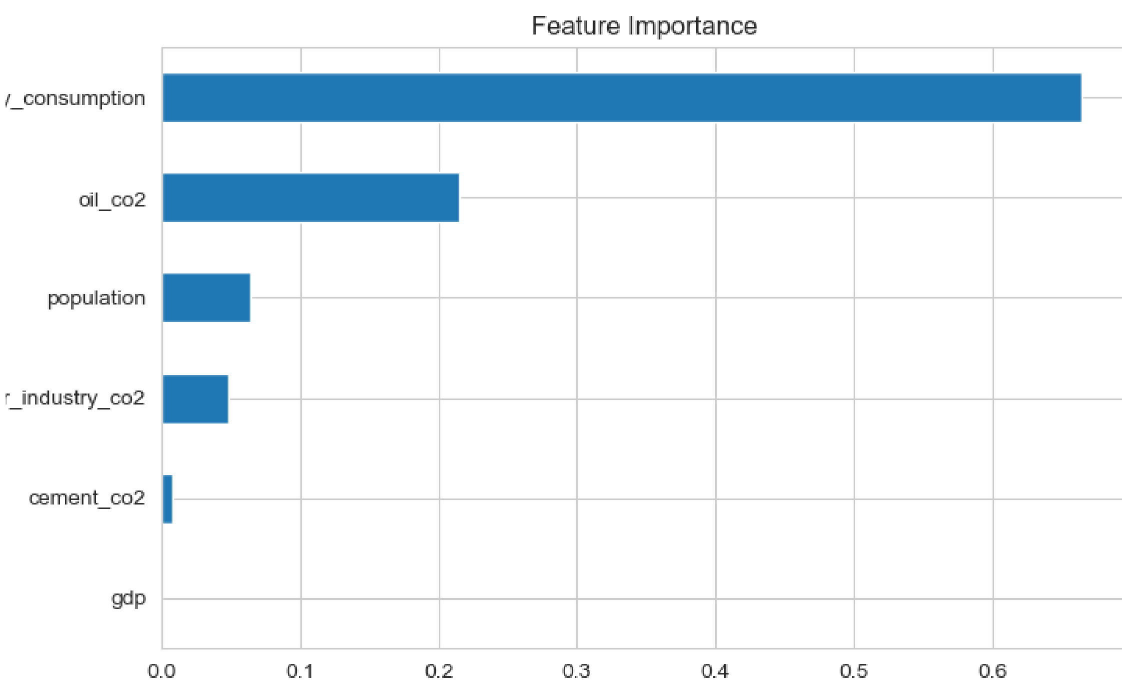

Figure 1: Feature Importance]) highlighted their impact. To enhance accuracy, feature engineering (e.g., energy_per_capita) and cross-validation were explored, with plans to test XGBoost for further improvement.

Figure 1: Feature importance plot from Random Forest, showing primary_energy_consumption (0.6648) as the dominant predictor, followed by oil_co2 (0.2147), population (0.0646), other_industry_co2 (0.0479), cement_co2 (0.0078), and gdp (0.0003).

Figure 2: Model comparison plot showing MAE and R^2 for Linear Regression, Random Forest, and (optionally) Tuned XGBoost, with Random Forest excelling.

Results

Feature importance analysis identified primary_energy_consumption (0.6648) as the strongest predictor, indicating energy use drives ~66.48% of CO2 emissions. oil_co2 (0.2147) underscored transport's role, while population (0.0646) and other_industry_co2 (0.0479) had moderate influence. Surprisingly, cement_co2 (0.0078) and gdp (0.0003) showed minimal impact, possibly due to data limitations or multicollinearity with energy consumption. Random Forest's high R^2 (0.8142) and lower MAE (332.04) suggest robust predictions, though the MAE's scale depends on whether co2 is in metric tons or millions.

Impact on SDG 13

These findings support SDG 13 by identifying energy consumption and transportation as critical targets for emission reduction. Policies promoting renewable energy and electric vehicles could leverage the model's insights, ensuring equitable resource allocation. For instance, regions with high primary_energy_consumption could prioritize green infrastructure, while transport-focused areas address oil_co2.

Ethical and Social Reflections

The model's reliance on energy data may bias predictions if industrial or economic factors (e.g., cement_co2, gdp) are underrepresented, especially in synthetic datasets. This could misguide policies in industrial-heavy regions. Ensuring data diversity and transparency in model outputs is crucial for fairness, particularly for low-income countries. Sustainability is enhanced by accurate targeting, but future models should incorporate renewable energy use to reflect green trends.

Conclusion

This project demonstrates machine learning's potential to address climate change, with Random Forest providing reliable CO_2 predictions. The emphasis on energy and transport offers actionable insights, though data quality and model enhancements (e.g., XGBoost) remain areas for growth. By refining the approach, we can better support global climate action.