Project Summary

SDG Problem Addressed

SDG 11.2 – Sustainable Cities and Communities:
Indicator 11.2.1 measures the *proportion of the urban population with convenient access to public transport* by sex, age, and disability status. Many cities lack consistent access to data. This project uses machine learning to predict and analyze access patterns across regions and years, helping identify where infrastructure improvements are needed.

ML Approach Used

Supervised Learning (Regression):
We trained a **Random Forest Regressor** using features such as:

- Country or Territory
- SDG Region & Sub-Region
- Data Reference Year

The model predicts the **share of the urban population with convenient public transport access (%).**

Workflow Summary

1. Load SDG 11.2 dataset (UN-Habitat XLS).
2. Preprocess → encode categorical → handle missing values.
3. Split data (80 % train / 20 % test).
4. Train **Random Forest Regressor**.
5. Evaluate using **MAE, RMSE, and R^2**.
6. Visualize feature importance and actual vs predicted values.
7. Save model (sdg11_2_model.pkl) for reuse or deployment.

Results

Metric	Value (example — insert your actual output)
MAE	~3.45
RMSE	~5.27
R^2 Score	~0.82

The model shows strong correlation between region, sub-region, and public-transport access, indicating potential policy insights for low-performing areas.

Ethical & Social Reflection

Bias Risks

- The dataset may under-represent smaller or developing cities, leading to biased predictions favoring data-rich regions.
- Historical inequalities (e.g., funding differences between continents) might be reflected in the data and unintentionally reinforced by the model.

Fairness & Sustainability

- By identifying underserved urban zones, the model supports *equitable access* to transport infrastructure.
- Encourages data-driven planning aligned with **sustainable, inclusive city development**.
- Transparency in methods (open-source data and interpretable features) promotes accountability.
