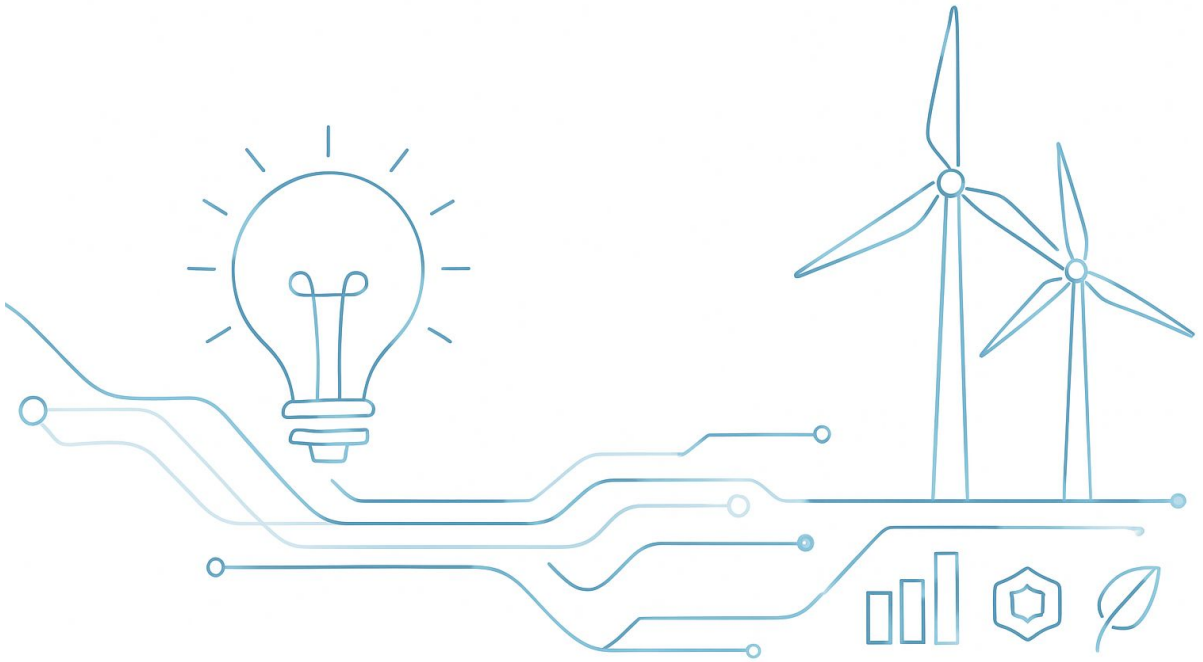
Machine Learning Approach to Renewable Energy Dynamics



1- Introduction

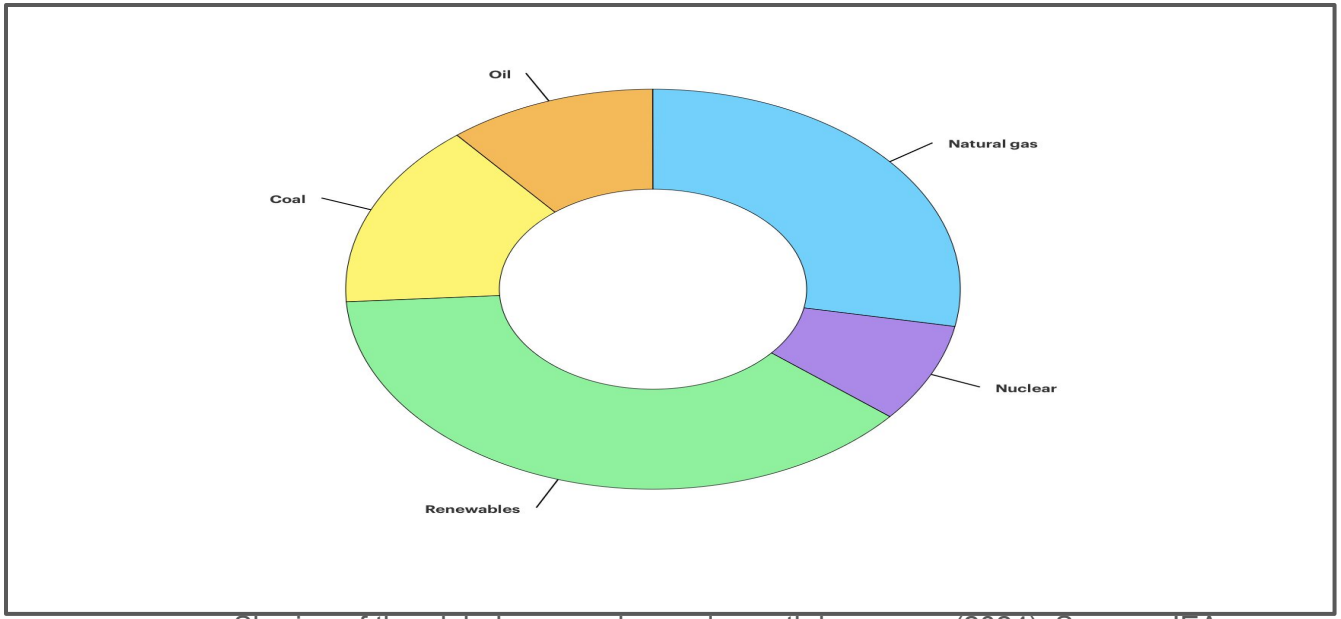
Context :

2024 -> The global energy demand increased by 2,2 %

- Much faster than the 10 years before means (1,3 %)
- Energy transition becoming more urgent.

Introduction

->Increased needs still mainly fueled by fossil fuels.



Sharing of the global energy demand growth by source (2024). Source : IEA

A fundamental question

Also, disparity between countries :

Iceland (81,03 %) and Somalia (95,03 %) vs Saudia Arabia (0,03 %) and Japon (7,69 %).

The "good students" and "late" countries are very different (wealth, climate, size...).

What are the features that determine the share of renewable energy in a country's final consumption ?

Key definitions

-Renewable energies:
Natural sources reconstituted faster than they are consumed.
Examples: solar, wind, hydraulic, tidal, geothermal, biomass.

-Final energy consumption:
Energy consumed directly by users (excludes energy used for production).

-Distinction:
Primary energy: crude, found in nature.
Secondary energy: transformed energy (e.g. electricity).

The Data

The Data show sustainable energy indicators and other useful factors across all countries from 2000 to 2020.

-The Variables: Y = Renewable energy share in total final energy consumption (%)
X = 3 types: Energy, Economic and Geographics

The features

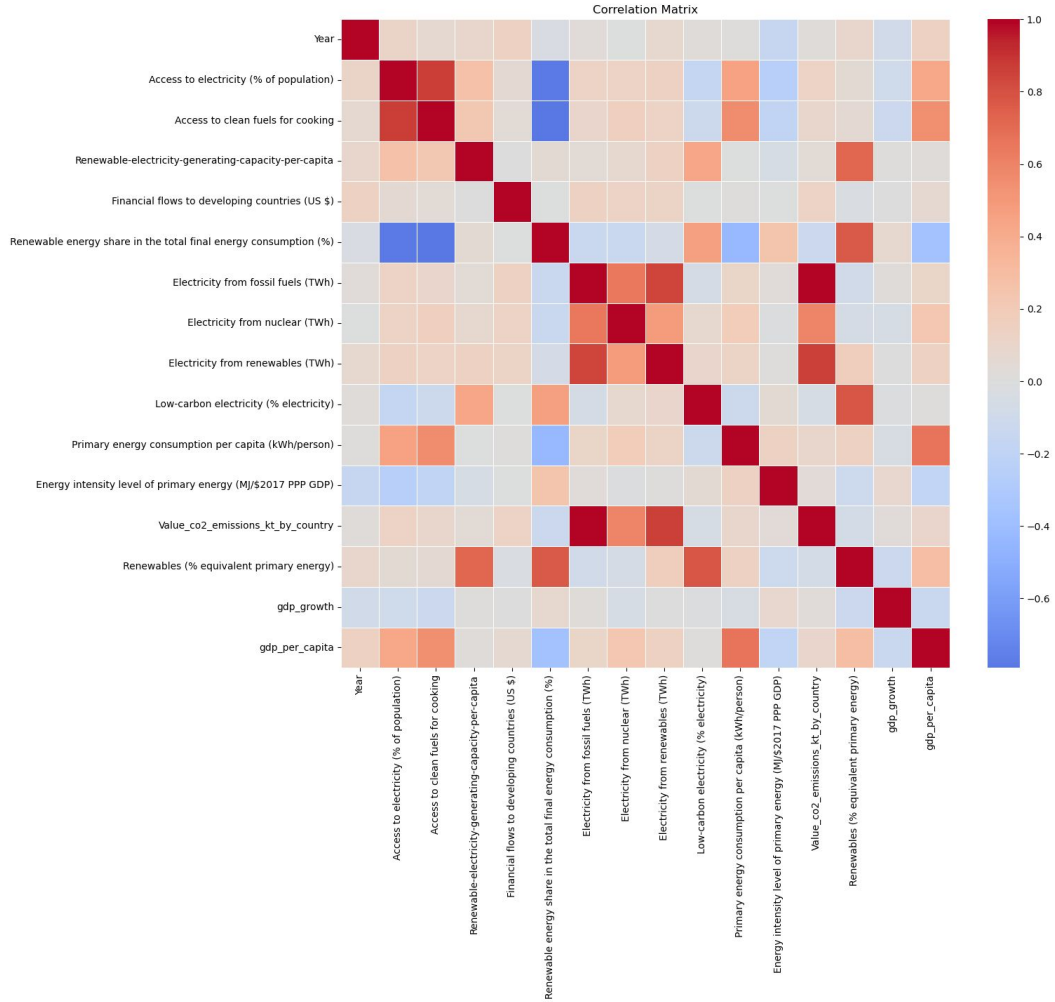
Energy :
Access to electricity (% population)
Access to clean fuels for cooking
Renewable production capacity per capita
Fossil, nuclear, renewable electricity (TWh)
Primary consumption per capita (kWh/person)
CO2 emissions per capita
Share of renewables in primary energy

Economy:
Financial flows to developing countries ($)
Energy intensity (MJ/$ GDP PPA 2011)
GDP per capita
GDP growth (%)

Demography & Geography:
Population density (P/km²)
Area (km²)
Latitude, Longitude

3- Data Preparation

- **Handling missing values:**
  - KNN Imputation with k=3
  - Standardization before and after imputation
  - Focused on preserving feature relationships
- **Restructuring the Dataset**
  - Country-year panel structure maintained
  - No aggregation to avoid losing year-by-year dynamics
  - Each line = a country at a given year
- **Feature Engineering**
  - Correlation Matrix analysis



Heatmap of the Correlation Matrix

- Removal of highly correlated variables:
  - Electricity from fossil fuels
  - Access to clean fuels for cooking

4- Methodology: Feature Selection

Features selected after mixed stepwise feature selection:

— Access to electricity (% of population)

— Renewable-electricity-generating-capacity-per-capita

— Primary energy consumption per capita (kWh/person)

— Energy intensity level of primary energy (MJ/$2017 PPP GDP)

— Renewables (% equivalent primary energy)

— GDP per capita

Methodology: Model Building

Four models were obtained:
— Linear Regression (Ordinary Least Squares)

— Ridge Regression (with cross-validation to select the optimal regularization parameter)

— Lasso Regression (with cross-validation)

— Random Forest Regression (non-linear ensemble method)

Methodology: Assessment methods

— Train/Test Performance : Models were evaluated based on R2 and Root Mean Squared Error (RMSE) on the test set.

— Cross-Validation : 5-fold cross-validation was used to assess model stability and generalization capacity.

— T-tests for Coefficient Significance : To determine a statistical significance of each feature for the OLS model, t-tests were performed. Features with p-values less than 0.05 were considered significant.

— Residual Diagnostics : Residuals were checked for normality and independence (Durbin-Watson test value around 1.92 indicated no severe autocorrelation)

Methodology: Stability of the model

Bootstrap analysis was performed:

— 1000 bootstrap samples were generated by resampling the training dataset with replacement.

— A Linear Regression model was fitted to each sample, and coefficients were recorded.

— Standard deviations of the bootstrapped coefficients were computed.

Methodology: Regularization Approach

— Ridge Regression (L2 penalty) was applied to shrink coefficients without setting any exactly to zero.

— Lasso Regression (L1 penalty) was applied, which can set some coefficients exactly to zero, which performed further feature selection.

— Random Forest Regression was applied to capture complex, non-linear relationships in the data

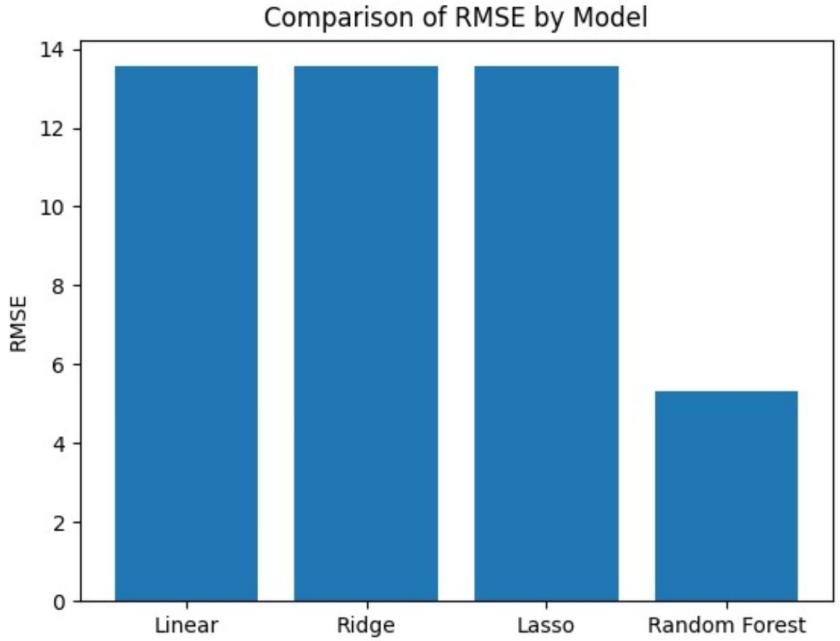
5- Results

Models Performance

Model	R^2 (Test Set)	RMSE (Test Set)
Linear Regression	0.7879	13.37
Ridge Regression	0.7881	13.36
Lasso Regression	0.7886	13.35
Random Forest	0.9839	3.69

- Linear, Ridge, and Lasso regressions show very similar performances ($R^2 \approx 0.788$, RMSE ≈ 13.35).
- Random Forest clearly outperforms the linear models ($R^2 \approx 0.984$, RMSE ≈ 3.69).
- This highlights the importance of capturing non-linear relationships between features and the target.

RMSE Comparison



- The barplot visually confirms that the Random Forest model achieves much lower prediction error compared to linear models.

- This suggests that complex interactions between variables are important for accurate renewable energy share prediction.

Cross- Validation

Model	Cross-Validation MSE
Linear Regression	229.24
Ridge Regression	229.25
Lasso Regression	229.24

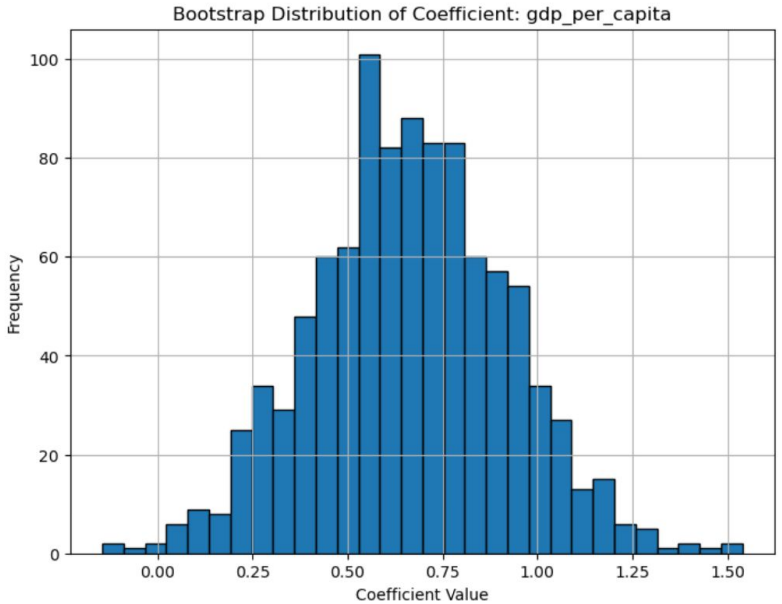
- 5-Fold Cross-Validation was performed for linear models.
- Cross-validation MSE values were very close (around 229), indicating stable model fitting.
- However, the test set performances remain limited compared to Random Forest.

Robustness (BootStrap)

- 1000 bootstrap samples generated
- Standard deviations of coefficients between 0.30 and 0.55
- Stability confirmed for key features (e.g., GDP per capita)
- Symmetric distribution observed for GDP per capita coefficient

Feature	Standard Deviation
Access to electricity (% of population)	0.475
Renewable electricity-generating capacity per capita	0.548
Primary energy consumption per capita (kWh/person)	0.411
Energy intensity level of primary energy (MJ/$GDP 2017 PPP)	0.477
Renewables (% equivalent primary energy)	0.499
GDP per capita	0.306

TABLE 4 – Bootstrap Standard Deviations of Selected Coefficients (1000 samples)



The Discussion

Missing data management: KNN imputation.

Conservation of annual variations to capture shocks and policies.

Access to electricity: linked to a decrease in the share of renewable energy.

GDP per capita: small but significant effect.

3 key factors: GDP per capita, primary consumption per capita, share of renewables.

Some counterintuitive interpretation :

Feature	Linear Regression	Ridge Regression	Lasso Regression
Access to electricity (% of population)	-22.19	-22.10	-22.19
Renewable electricity generating capacity per capita	4.87	4.81	4.80
Primary energy consumption per capita (kWh/person)	-3.35	-3.33	-3.26
Energy intensity level of primary energy (MJ/$2017 PPP GDP)	3.14	3.10	3.03
Renewables (% equivalent primary energy)	9.40	9.42	9.37
GDP per capita	-0.78	-0.80	-0.77

Limitations

- Changing the model, or maybe manipulating the data better?

- Developing countries vs rich countries?

- Data engineering: adding more data?