Machine Learning Approach to Renewable Energy Dynamics

Presented by
Grzegorz Mozdzynski
Hajar Sriri
Chopard Jules

University of Bicocca

28 avril 2025

Table des matières

1 Introduction

1.1 Context

According to the IEA (International Energy Agency), in 2024 the global energy demand rose by 2,2 % compared to 2023. In comparison, it's considerably faster than the average annual demand increase of 1.3% between 2013 and 2023.

-This rise in energy needs makes the transition towards renewable sources of energy even more imperative and drive country to accelerate their energy transition. However, it is important to note that the current increase in global energy needs is still largely fueled by fossil sources such as coal and oil 1. Against this backdrop, a critical observation emerges : there is a significant disparity between countries regarding the share of renewable energy in their final energy consumption.

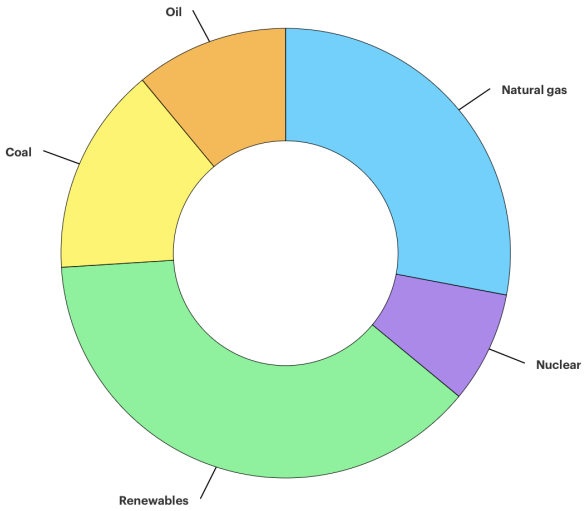

FIGURE 1 – Share of the global energy demand growth by source (2024). Source : International Energy Agency

For exemple in 2019 (81,07%) of the final energy consumed in Island is from renewable energy sources and this share is higher in Somalia with (95,03%). So those countries could be considered as "good students" in the global energy transition challenge. On an other hand some countries like Saudia Arabia (0,03%) or Japan(7,69 %) still for 2019, continue to use a high share of fossile source on their national final energy consumption.

Through this observation, we can constat one particular thing. Indeed, we can easily observe that regarding the "good students" of the energy challenge. The 2 countries we took in example (Iceland and Somalia) are completely different (wealth, climate, demography, size ...) and it's exactly the same constat for the countries with low renewable energy share (Japon and Saudia Arabia) they are totally different. This fact drive us to a fundamental and interesting question which the following one :**What are the feature**

which determine the share of renewable energy in a country's final energy consumption ?

1.2 Definitions

Before we start it's important to give a definition of the main term we will approach in our analysis. First of all we call "renewable energy" all the energies derived from natural sources that are replenished at a higher rate than they are consumed. There are 7 types which are :

- Solar energy.
- Wind energy.
- Hydro energy.
- Tidal energy.
- Geothermal energy.
- Biomass energy.

According to the European Commission, the final energy consumption of a country is the total amount of energy consumed by end users, such as households, industry and agriculture. It is the energy which reaches the final consumer's door and excludes that which is used by the energy sector itself. For example, the gasoline burned in a car is a final consumption, or the electricity using for light or industrial machine is a final consumption. But the coal used to produce the electricity who's going to be used for our industrial machine is not the final consumption.

In summary the renewable energy share in total final energy consumption is the part of energy consumed by end user coming from renewable source.

Finally it's important to know a distinction between primary energy and secondary energy. Basically, primary energy is refering to the energy found in nature that has not been subjected to any human engineered conversion process. It encompasses energy contained in raw fuels and other forms of energy, including waste, received as input to a system. Primary energy can be non-renewable or renewable. One the other side, we have the Secondary energy which is a carrier of energy, such as electricity. These are produced by conversion from a primary energy source.

2 Data description

2.1 The data set

Our data set is from a data set library "Kaggle" and it's gathered data from the World Bank, the International Energy Agency (IEA) and the website "Our world in Data". The

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

2.2 The variables

Now let's talk about the variables composing the data set. First of all we have our described variable which is as we said it before, the **Renewable energy share in total final energy consumption (%)**.

The describing variable can be divided into differents groups :

- **The Energy variables :**
  - *Access to electricity (% of population).*
  - *Access to clean fuels for cooking.*
  - *Renewable electricity generating capacity per capita.*
  - *Electricity from fossil fuels (TWh)*
  - *Electricity from nuclear (TWh)*
  - *Electricity from renewables (TWh)*
  - *Low-carbon electricity (% electricity)*
  - *Primary energy consumption per capita (kWh/person)*
  - *Value co2 emissions (metric tons per capita)*
  - *Renewables (% equivalent primary energy)*

- **The Economic variables :**
  - *Financial flows to developing countries (US $)*
  - *Energy intensity level of primary energy (MJ/$2011 PPP GDP)*
  - *GDP per capita*
  - *GDP growth (annual %)*

- **The Demographic and Geographic variables :**
  - *Density (P/Km2)*
  - *Land Area (Km2).*
  - *Latitude.*
  - *Longitude*

3 Data preparation

3.1 Handling Missing Values

- Removal of countries with insufficient data for key features.

We faced several variables that exhibited missing values due to gaps in data collection for certain countries and years. Indeed, in the original sustainable energy dataset, missing values were frequent across several key features, such GDP growth, primary energy consumption per capita among others. Rather than removing all records with missing values which would have led to a lake of data we chose to impute them.

- KNN imputation of missing values within each country-year group.

We used a K-nearest Neighbors (KNN) Imputer to fill missing values with k=3. Why k=3 ? We opted for a small number of neighbors to prioritize local similarity and a smaller k prevents over smoothing the data. KNN Imputation was preferred over mean imputation to better preserve the trend of the data and the relationships between variables. Before applying KNN, the features were standardized to ensure that distance calculations were meaningful and not dominated by features with larger scales like GDP per capita. After the KNN imputation a second standardization was performed to maintain normalized feature distributions, which is essential for the stability and comparability of machine learning models such as Linear Regression, Lasso and Ridge Regression.

— Analysis by country-year instead of country averages.

Each observation in our dataset corresponds to a specific country-year combination, based on the panel structure of the original data.

Rather than aggregating the data by country (e.g., by computing country averages), we chose to preserve the year-by-year dynamics. This decision was made to capture short-term variations and trends in renewable energy usage, which could be influenced by yearly policy changes, economic developments, or external shocks such as financial crises or global agreements on climate change.

Aggregating the data would have masked these critical temporal dynamics and potentially reduced the predictive power of our models. By maintaining the panel structure, our analysis is better suited to reflect the evolving nature of sustainable energy adoption over time.

3.2 Features Engineering

— Correlation matrix analysis

In order to better understand the relationships between our features and to avoid issues related to multicollinearity we computed and visualized the correlation matrix for all numeric variables. This matrix shows what it called the Pearson correlation coefficient between each pair of variables. It can be interpreted like that : -1 is a perfect negative correlation to +1 is a perfect positive correlation. Having strong correlations can introduce redundancy and lead to instability in linear models such as OLS. A heatmap was created to visually inspect these correlations, there is the figure below.

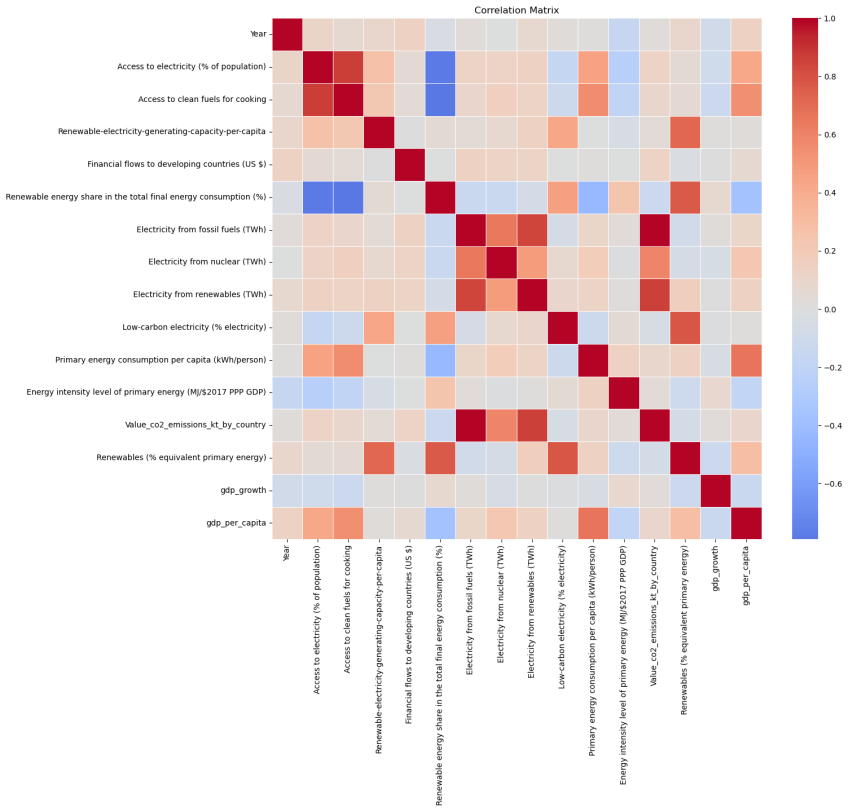

FIGURE 2 – Correlation Matrix

— Removal of highly correlated predictors to avoid multicollinearity.

Based on the Correlation Matrix it was decided to remove fuel electricity from fossil fuels access to clean fuel for cooking because too highly correlated based on the matrix. Indeed these features were strongly correlated with other variables such as access to electricity and low carbon electricity generation, which make them redundant. This step is very important and crucial in supervised learning, because it ensures that the models will not be affected by multicollinearity which can inflate the variance and distort the coefficient estimates.

4 Methodology

In this section, we present the methodology used to construct, evaluate, and validate our predictive models for renewable energy share.

4.1 Feature Selection

First topic to investigate is feature selection. We deleted correlated features based on correlation matrix, but we still have to select final features that will be included in the model. For that we used **mixed stepwise feature selection**, which is a combination of forward and backward selection, using Sequential Floating Selection and 5-fold cross-validation. The selection criterion was based on maximizing the R^2 score. As a result, six features were selected :

— 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

4.2 Model Building

After these steps we are working with clean, ready to use data and we have an optimal set of features, so we can proceed with model building. We chose different models to investigate different methods for predicting Renewables Share.

— 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)

The dataset was split into 80% for training and 20% for testing. All linear models were trained on standardized features to improve convergence and stability.

4.3 Assessment Methods

Model evaluation was performed using several strategies :

— **Train/Test Performance :** Models were evaluated based on R^2 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).

4.4 Stability of the Model

To verify the stability of the linear regression model coefficients, we performed a **bootstrap analysis** :

— 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.

The standard deviations ranged between approximately 0.30 and 0.55, confirming good stability of the estimated coefficients.

4.5 Regularization Approach

To further address potential multicollinearity and improve the model :

— 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.

While Ridge and Lasso regressions provided minor improvements in performance compared to standard Linear Regression, Random Forest Regression significantly outperformed all linear models in terms of predictive accuracy, confirming the presence of complex, non-linear relationships in the data. However the results can suggest overfitting, which sometimes takes place, especially for tree-based methods.

5 Results

In this section, we present the outcomes of our feature selection process and evaluate the performance of several predictive models developed to forecast the share of renewable energy consumption across countries and years. We first summarize the variables retained after data preparation, then assess the performance of linear models (Linear Regression, Ridge, and Lasso) alongside a non-linear Random Forest model. Comparisons are based on test set errors, cross-validation metrics, and model interpretability through coefficient analysis.

5.1 Summary of feature selection outcomes

— Which predictors were retained

Thanks to the correlation matrix analysis we removed "Electricity from fossil fuels (TWh)" and "Access to clean fuels for cooking" to avoid multicollinearity and biais.

— T-tests for coefficient significance

The statistical significance of every feature was determined by performing t-test for each feature. The table below summarizes the results :

Feature	**t-statistic**	**p-value**
Access to electricity (% of population)	-60.641	<0.001
Renewable-electricity-generating-capacity-per-capita	17.726	<0.001
Primary energy consumption per capita (kWh/person)	-7.757	<0.001
Energy intensity level of primary energy (MJ/$2017 PPP GDP)	9.519	<0.001
Renewables (% equivalent primary energy)	26.529	<0.001
GDP per capita	-1.937	0.053

TABLE 1 – T-tests for feature coefficients (OLS)

After analysing the Table 1, almost all of the features were determined to be statistically significant at the 5% level (p-value < 0.05). Only one : *GDP per capita* was non-significant, but the p-value ≈ 0.053 indicates that it is only marginally non-significant.

5.2 Performance of the Models

— Train and validation error (MSE) for linear, Ridge, and Lasso models.

We fitted four models to predict the share of renewable energy : Linear Regression Ridge Regression Lasso Regression Random Forest Regressor

The performances on the test set are summarized below :

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

TABLE 2 – Test Set Performances for Different Models

Additionally, 5-fold cross-validation was performed on the training set for linear models :

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

TABLE 3 – Cross-Validation Mean Squared Errors (5-Fold CV)

The Cross-validation results confirm that the linear models behave consistently across different folds, indicating stable fitting.

— Comparison plots (e.g., MSE vs model type).

To visually compare model performances, we plotted the RMSE for each model :

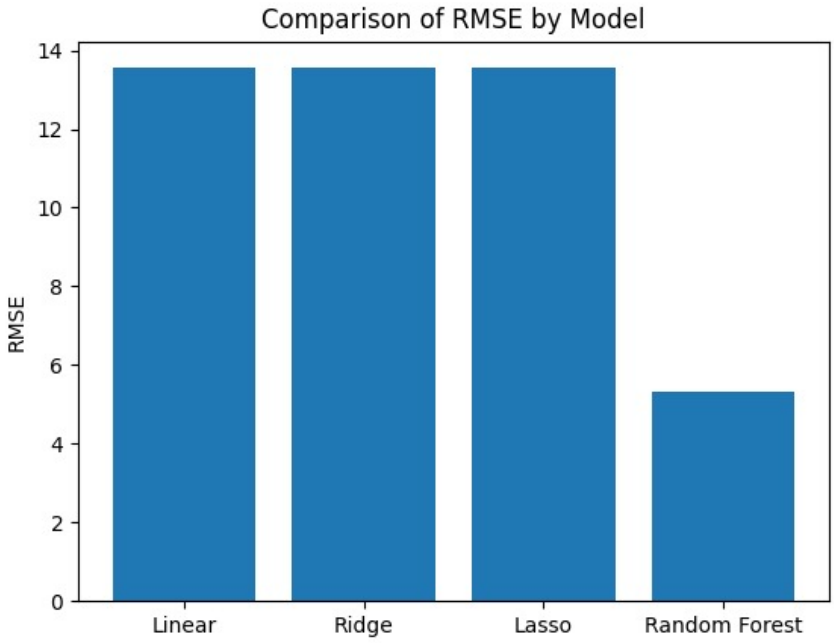

FIGURE 3 – Comparaison of RMSE by models

From the plot, it is clear that Random Forest Significantly outperforms all linear models, achieving a much lower RMSE compared to any linear model. This suggest that non-linear relationships and complex feature interactions are crucial to modeling the share of renewable energy consumption across countries and years.

5.3 Coefficient Analysis and Robustness

— Discussion on significant variables

Analyzing the coefficients obtained from the linea models provides future insights : Indeed, the Linear Regression and the Ridge Regression showed that features such as GDP per capita, Energy intensity and access to electricity had significant impacts on renewable energy share. Ridge Regression applied uniform shrinkage to all coefficients, reducing their magnitude but preserving all predictors in the model. Lasso Regression slightly reduce the influence of certain less important predictors but did not set any coefficient exactly to zero, indicating that all selected features contributed for information. These results show that the features kept after pre-cleaning were already useful, and that feature selection removed any remaining irrelevant predictors.

— Sensitivity of estimates under bootstrap.

To assess the robustness of our linear regression model coefficients, we applied a bootstrap resampling procedure.

As shown in Table 4, the standard deviations of the estimated coefficients range between approximately 0.30 and 0.55. These relatively low values indicate that the estimated effects of the features are stable across different resamplings of the training data.

In particular, features such as *GDP per capita*, *Primary energy consumption per capita*, and *Renewables (% equivalent primary energy)* exhibit consistently robust influence on the renewable energy share prediction.

Thus, the bootstrap analysis reinforces the reliability of the selected predictors and validates the stability of the linear regression model under moderate sampling variability.

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)

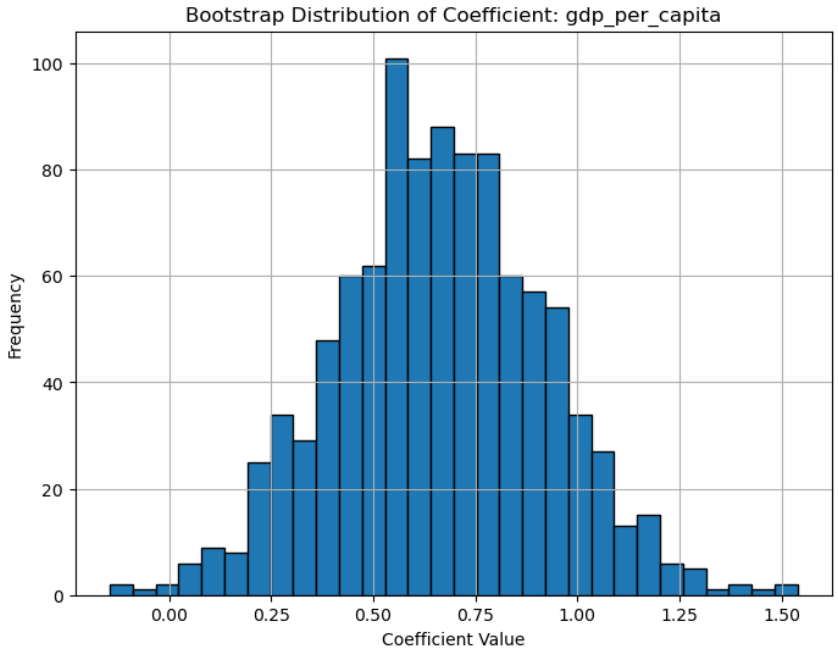

FIGURE 4 – Bootstrap distribution of the coefficient for *GDP per capita*.

Figure 4 displays the distribution of the estimated coefficient for *GDP per capita* across 1000 bootstrap samples.

The distribution is approximately symmetric and bell-shaped, centered around 0.65, with most values ranging between 0.25 and 1.00. This pattern suggests that the influence of GDP per capita on the renewable energy share is consistently positive and relatively stable across different data resamplings. The narrow spread of the distribution further confirms the robustness of this feature's estimated effect.

6 Discussions and limits

6.1 Discussion

In our study we were asking to explain the different element explaining the share of Renewable Energy on the final energy consumption for each country. Basically our data

set was composed by missings values that we handled using a KNN imputation in order to preserve the trend of the data and the relationship between variables. Also we voluntary decided to let the years-by-years dynamics in order as we said to capture short term variation and then policy change, or external shock.

Then we made a first selection of variables by excuding the one too correlated with the others. Then we did a second selection of the most statistically relevant variables using a mixed stepwise feature selection.

After cleaned our data we started to build our model using different prediction methods 3 linears (Linear regression, Lasso regression and Ridge regression) and one non-linear approach the Random forest regression.

Then, we used different methods to assess the performance and the stability of our models.

Then to return to our question, we have several things to say. First we can say that all of our 6 factors variables are relatively statistically significant except for the GDP/cap but it's marginally. Second, our results show that, although linear models offer valuable interpretability and good robustness, it was the Random Forest model that achieved the best predictive performance, better capturing complex and non-linear relationships present in the data.

Also among our features, 3 of them can be highlight for their particular robustness and consistent influence for the prediction of the RE share on final energy consumption. The feature are the GDP-CAP, the primary energy consumption per capita and the renewable (so the equivalent primary energy that is derived from renewable sources.)

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

TABLE 5 – Comparison of Feature Coefficients across Linear, Ridge, and Lasso Regressions

To come back to our question as we can see in Figure 5 our results suggest that country with a better access to electricity have also a lower share of RE on their final energy consumption, this could be contreintuitive but in fact in developping countries the increase of the access to electricity is not always fueled by RE but by fossile energy.

Finaly our analysis revealed that GDP per capita, while statistically significant, had only a marginal impact on the renewable energy share, suggesting that economic factors might play a secondary role compared to factors directly related to energy access and renewable infrastructure.

6.2 limits

One of the biggest factors limiting the performance of the models could be the data itself. Unfortunately we had to work with a lot of missing values. We tested two different imputation techniques : mean and KNN with k=3. Mean refleating flattened variability and weakened the signal resulting in lower Lasso performance (RMSE=12,03). So using KNN imputation, preserved features relationship and improved model accuracy. However better imputation techniques may be reasonable to investigate, for example imputing the missing value not only based on the nearest neighbour, but also based on the country's

economical state, geographical location. Further investigation of the data preparation techniques may positively impact the models performance.