# Topic: Evolution of Renewable Energy Share in the Energy Mix Across Countries Machine Learning

L. Seels, C. Soulard, M. Tinel

Machine Learning – ESILV, EVD-3

## ABSTRACT

*This project applies machine-learning techniques to analyze and forecast the evolution of renewable-energy usage worldwide. Using a Kaggle dataset containing historical renewable-energy shares for each country, combined with World Bank income-group information, we build a database suitable for statistical analysis and modeling. After exploring the global trend, we evaluated several models including Linear Regression, Decision Trees, k-NN, and SVM using standard performance metrics and hyperparameter tuning. The RBF-kernel SVM delivers the best accuracy and is used to generate projections up to 2050 both globally and for each income group. The results reveal strong disparities across income levels, with high-income countries showing rapid growth while lower-income groups progress more slowly. When compared with international climate objectives, the projected renewable shares remain insufficient to meet the thresholds required to limit global warming to* 1.5 °C *or even* 2 °C*. Overall, the project demonstrates how data analysis and machine learning can be used to study the energy transition and evaluate its alignment with climate goals.*

**Key words.** renewable energy – machine learning – regression – SVM – energy transition

## Contents

## 1. Introduction

In this project we want to study the evolution of the share of renewable energy in the energy mix and verify if the trend follows the global objective. To do so we applied machine-learning methods to analyze and forecast the evolution of renewable-energy usage worldwide. We used a Kaggle dataset that reports the renewable share of primary energy in their energy mix for each country and, by enriching it with World Bank income-group information, we created a clean dataset that allows us to study both global trends and differences between economic groups.

We first imported and cleaned the Kaggle dataset on renewable-energy shares, keeping only valid countries and the relevant columns. We then explored the data through basic statistics and visualizations to understand the global trend. After formalizing the task as a regression problem, we trained several models (Linear Regression, Decision Tree, k-NN, and SVM) using a randomized train/test split and evaluated them with MAE, RMSE, and $R^2$. With hyperparameter tuning, the SVM emerged as the best-performing model, and we used it to produce a global projection up to 2050.

In the second part, we enriched the dataset with World Bank income groups and built a clean dataframe containing each country, renewable share, and income category for every years. We computed the average trend per group, plotted the historical evolution, and trained a SVM model for each income group to generate future trajectories. Finally, we analyzed these results and compared the projected growth with international climate goals.

Overall, the project illustrates how machine-learning models can be used to explore the energy transition and assess the alignment between observed trends and long-term climate targets.

## 2. Defining the problem and searching for data

In this first step, we start by clearly defining the problem we want to solve and identifying the type of data required to address it. Our objective is to:

- analyze how the share of renewable energy evolves over time across different countries;
- build machine-learning models capable of predicting future trends.

Because the energy transition is closely linked to a country's level of development, we also aim to compare these trends across income groups.

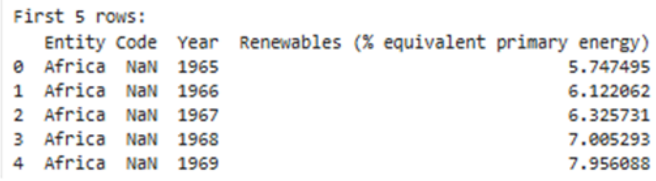
Once the problem is defined, we search for reliable datasets that contain long-term historical values. For this project, we use a Kaggle dataset providing the renewable-energy share for each country over several decades, and we complement it with the World Bank "Country and Income Group" dataset to classify countries by level of income.

The first dataset is called *"Renewable Energy Worldwide: 1965–2022"*, and it is available on Kaggle:

https://www.kaggle.com/datasets/belayethossainds/renewable-energy-world-wide-19652022

This dataset contains data that gives an overview of the global transition toward sustainable energy, compiling country-level indicators that reflect progress in reducing reliance on fossil fuels and adopting cleaner energy sources. It includes variables such as access to electricity, availability of clean cooking fuels, renewable energy generation per capita, shares of renewable and low-carbon electricity, etc. By combining environmental, social, and economic dimensions, the dataset enables researchers, policymakers, and analysts to explore patterns and trends in how different nations are advancing toward sustainable, low-carbon energy systems, supporting data-driven decisions for global climate and energy policy.

For our work we will study the dataset about the share of renewable energy in the energy mix of different countries and regions called *"renewable-share-energy"*. This dataset provides a broad view of the global adoption of renewable energy, covering country-level metrics over time on how nations are deploying clean energy technologies, increasing the share of renewables in their energy mix, and transforming their energy infrastructure for a sustainable future.



```
First 5 rows:
  Entity Code  Year  Renewables (% equivalent primary energy)
0 Africa  NaN  1965                                  5.747495
1 Africa  NaN  1966                                  6.122062
2 Africa  NaN  1967                                  6.325731
3 Africa  NaN  1968                                  7.005293
4 Africa  NaN  1969                                  7.956088
```

Figure 1: Renewable share energy dataset

The second dataset about the income group is also available on Kaggle:

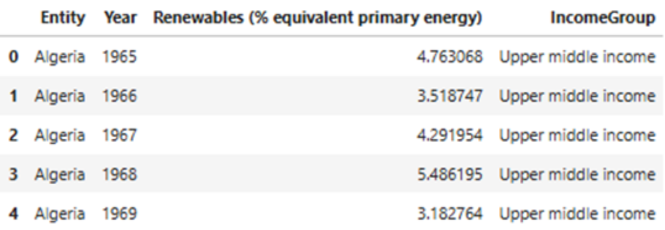
https://www.kaggle.com/datasets/taniaj/world-bank-country-and-lending-groups?select=worldbank_classification.csv

This second dataset provides country-level economic classification based on the World Bank's income grouping system. It assigns each country to categories such as High income, Upper middle income, Lower middle income, or Low income, using standardized criteria grounded in Gross National Income (GNI) per capita. This classification reflects essential socio-economic differences between nations, helping analysts understand how economic development shapes access to resources, technological capacity, and investment potential. By offering a simple yet powerful way to compare countries according to their level of income, the dataset serves as an important reference for studies in global development, economics, and public policy.

In our project, this dataset is used to enrich the renewable-energy dataset by linking each country to its corresponding income group. This allows us to analyze renewable-energy adoption not only at the global level but also across different economic categories. By integrating this socio-economic dimension into our analysis, we can explore how income level influences renewable-energy growth, identify disparities between regions, and compare long-term projections across income groups. This enables a deeper understanding of the global energy transition through the combined lens of technological progress and economic development.

Together, these two sources give us the necessary information to perform both global analysis and economic comparisons.

In this merged dataset we can see the country, the years, the corresponding renewables share, and the income group.

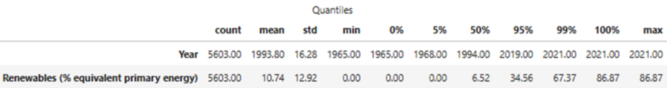


|  | Entity | Year | Renewables (% equivalent primary energy) | IncomeGroup |
|---|---|---|---|---|
| 0 | Algeria | 1965 | 4.763068 | Upper middle income |
| 1 | Algeria | 1966 | 3.518747 | Upper middle income |
| 2 | Algeria | 1967 | 4.291954 | Upper middle income |
| 3 | Algeria | 1968 | 5.486195 | Upper middle income |
| 4 | Algeria | 1969 | 3.182764 | Upper middle income |

Figure 2: Merged dataset

## 3. Cleaning and exploring the dataset

In this step, we focus on preparing the data so it can be reliably used for analysis and modeling. We removed the `Code` column that was not useful for our study, and we checked for missing or absurd values. Once the dataset is cleaned, we explore it through descriptive statistics and visualizations to identify general trends, variations across countries, and potential anomalies.



|  | Quantiles |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  | count | mean | std | min | 0% | 5% | 50% | 95% | 99% | 100% | max |
| Year | 5603.00 | 1993.80 | 16.28 | 1965.00 | 1965.00 | 1968.00 | 1994.00 | 2019.00 | 2021.00 | 2021.00 | 2021.00 |
| Renewables (% equivalent primary energy) | 5603.00 | 10.74 | 12.92 | 0.00 | 0.00 | 0.00 | 6.52 | 34.56 | 67.37 | 86.87 | 86.87 |

Figure 3: Global information about our dataset

Here we can see that the minimal value for the share of renewable energy in the mix of a country is 0% and the highest is 86.87%. In this first graph we can see that the global renewable energy share grows over time. It went from 6% in 1975 to 13% in 2020. We can also observe a sharp acceleration in the last years, caused by global politics for ecological transition.

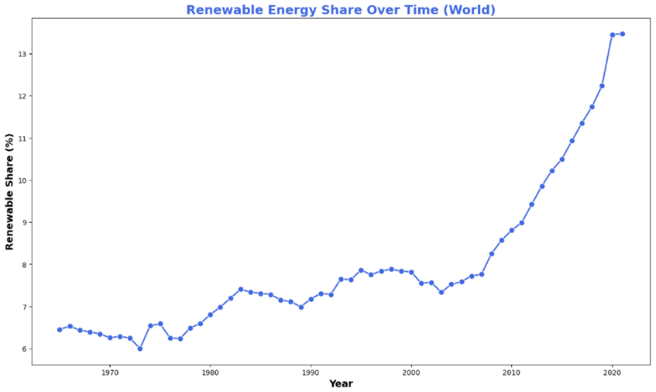
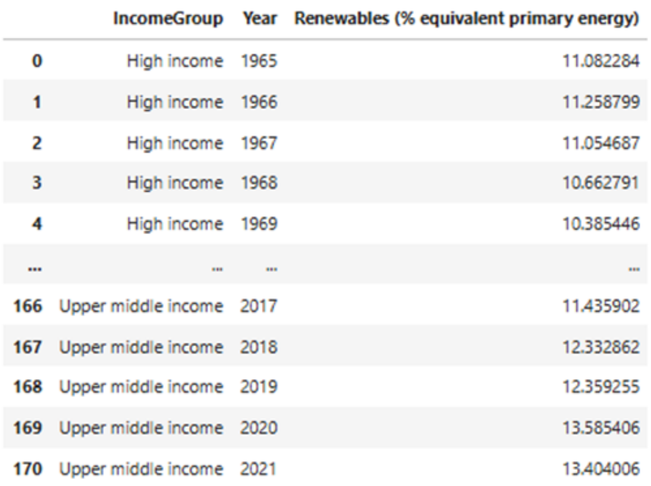


Figure 4: Renewable energy share over time for the world

Plotting the historical evolution of renewable-energy shares allows us to observe global progress over time, while comparing distributions across income groups helps reveal structural differences in the pace of the energy transition. This exploratory phase provides the foundation for defining meaningful features, understanding the behavior of the data, and guiding the choice of appropriate machine-learning models for forecasting.

To explore the difference between every country's income group we merged our two datasets. We began by removing aggregated entries such as continents or "World", keeping only individual countries with valid numerical values. We then averaged the values for every income group, and we obtained the dataset and graph below.

This graph shows how renewable-energy usage evolves across income groups. This first graph already reveals important differences: high-income countries generally show higher renewable growth in recent years, while lower-income categories evolve more slowly. This visualization provides a direct, intuitive overview of the disparity between income groups.



| | IncomeGroup | Year | Renewables (% equivalent primary energy) |
|---|---|---|---|
| 0 | High income | 1965 | 11.082284 |
| 1 | High income | 1966 | 11.258799 |
| 2 | High income | 1967 | 11.054687 |
| 3 | High income | 1968 | 10.662791 |
| 4 | High income | 1969 | 10.385446 |
| ... | ... | ... | ... |
| 166 | Upper middle income | 2017 | 11.435902 |
| 167 | Upper middle income | 2018 | 12.332862 |
| 168 | Upper middle income | 2019 | 12.359255 |
| 169 | Upper middle income | 2020 | 13.585406 |
| 170 | Upper middle income | 2021 | 13.404006 |

Figure 5: Renewable energy in the energy mix by income group

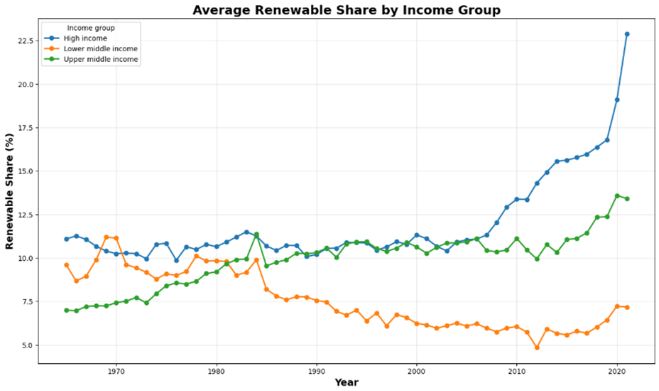


Figure 6: Renewable energy share by income group ).

This representation allow us to compare trend without the noise of individual national fluctuations. The figure makes the structural differences between groups immediately visible: high-income countries generally exhibit the strongest growth in renewable-energy share, particularly in recent years, while upper-middle-income countries progress at a steadier but more moderate pace. In contrast, lower-middle-income economies show much slower improvements, resulting in a widening gap between them and higher-income groups. Overall, this visualization provides a clear and intuitive overview of how economic development influences the pace of the energy transition across the world.

## 4. Analysis using models

We applied several models to forecast the future evolution of renewable energy share.

- **k-Nearest Neighbors (k-NN)** predicts values based on the closest historical points, which allows it to fit past data very closely but limits its ability to extrapolate beyond the observed range.
- **Decision Trees** work by splitting the timeline into intervals and assigning a constant value to each segment based on the training data. This can closely follow historical trends, but predictions are constant and often fail to capture smooth, long-term growth.
- **Linear Regression** fits a straight line through the data, providing a simple global trend. While it captures the overall direction of change, it cannot adapt to nonlinear accelerations or decelerations in the series, which may lead to under or overestimations in the long term.
- **Support Vector Machines (SVM)** using an RBF kernel, model a smooth nonlinear trend that extends the observed dynamics more plausibly into the future. Although SVM may not always outperform other models on standard metrics, it is generally the best at capturing long-term trends.

We first started by using the linear regression, and we uplifted the regression to match the last point of historical data.

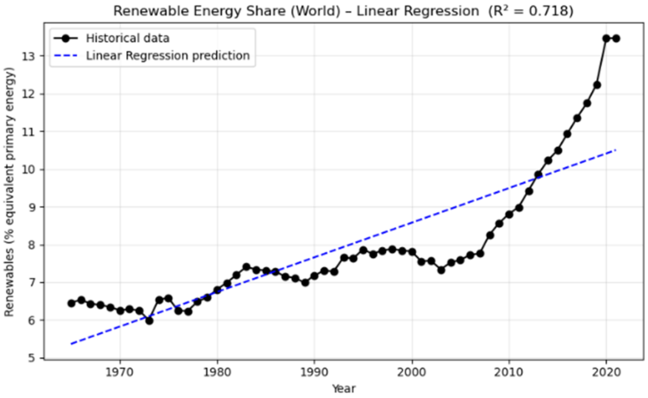


Figure 7: Linear regression model .

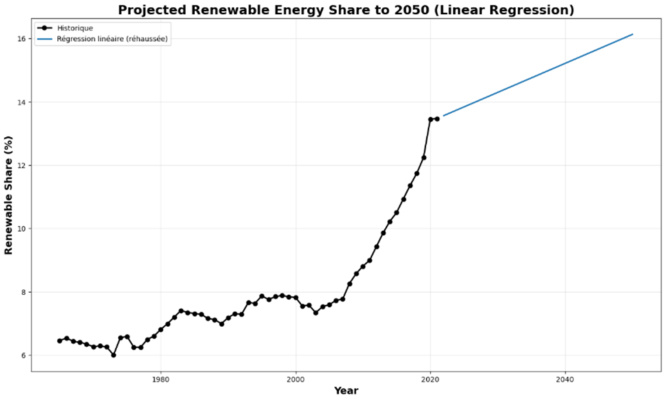


Figure 8: Projection to 2050 by linear regression

We can see with the linear regression that the predicted values follow an average growth trend based on our dataset. However, this may not be an appropriate model because the growth in recent years has been increasing very quickly.

We then used the SVM. We first split our data in two, train and test, and applied a shuffle on the train/test split to avoid training data only on the first years. This avoids the training being done only on the first years, so the model also sees the accelerated growth in the last years.

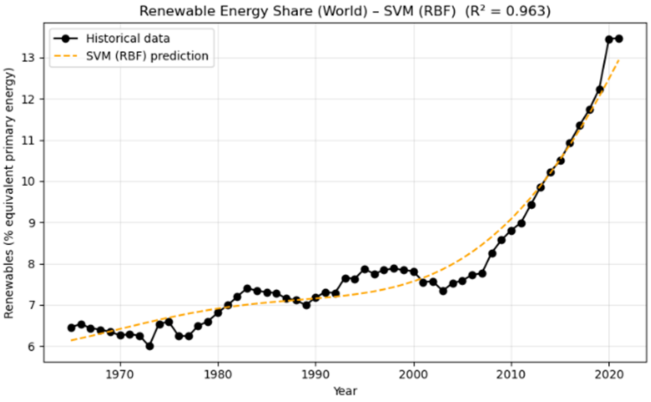


Figure 9: SVM model

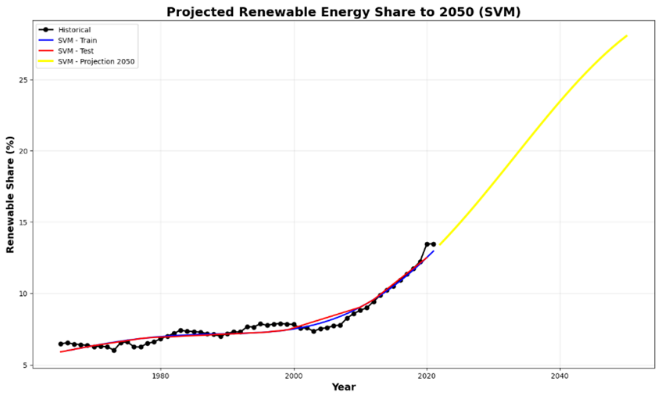


Figure 10: SVM and projection to 2050, with train (blue) and test (red) data

From the SVM projection, we can see a smooth and quickly rising curve right after the most recent historical values. Since the RBF kernel captures nonlinear patterns, the model extends the recent rapid growth into the future, resulting in a trajectory that increases faster than a simple linear model. This suggests a continued and significant rise in renewable energy share in the coming decades, consistent with the strong acceleration observed in recent years. However, as with any nonlinear extrapolation, these values should be interpreted with caution because the model assumes that the recent rapid growth continues at a similar pace.

We also tested two additional models, k-NN and Decision Trees, to evaluate their ability to forecast the future evolution of renewable-energy share. In the case of k-NN, predictions are made by averaging the values of the closest years in the training set. This approach works reasonably well for interpolation within the observed range, but it is fundamentally unsuited for long-term forecasting. When predicting future years, the model has no neighbors beyond the last available data, so it repeatedly selects the same past points as nearest neighbors. As a result, the projections give a flat line that cannot represent the accelerating growth observed in recent years.

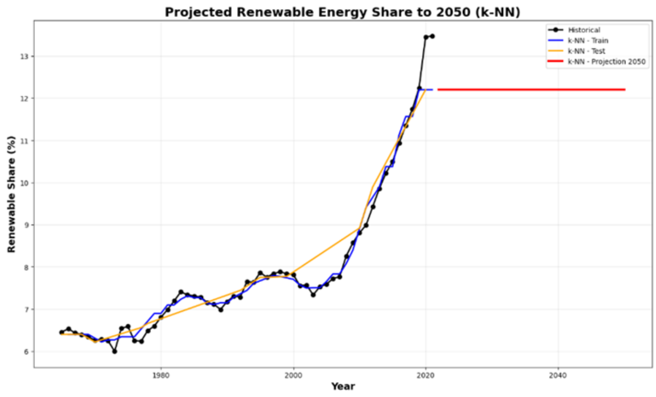


Figure 11: k-NN model and projection to 2050

Similarly, the Decision Tree model faces a different but related limitation. By construction, it divides the timeline into segments and assigns a constant value to each one. This allows the tree to fit the historical data closely, but it produces constant predictions that cannot follow a smooth upward trajectory. Instead of continuing the recent acceleration in renewable energy adoption, the model generates a flat level, which does not correspond to realistic long-term behavior.

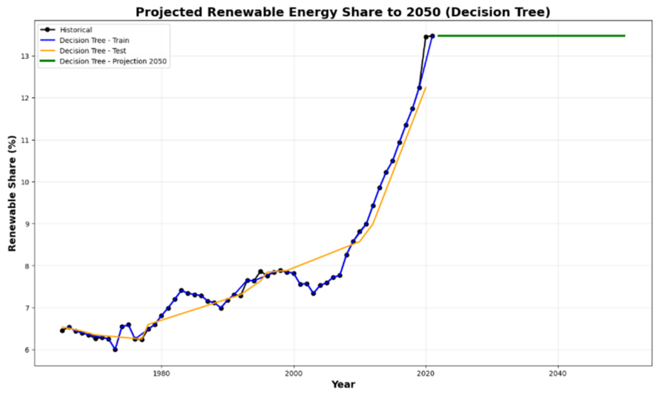


Figure 12: Decision tree model and projection to 2050

For these reasons, both k-NN and Decision Trees are mainly included for comparison in our analysis. They help illustrate the limitations of simpler or more rigid methods, while highlighting the need for more flexible models such as SVM that can capture non-linear dynamics and provide more credible long-term forecasts.

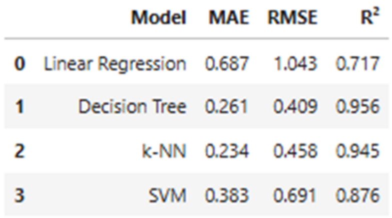
To conclude by comparing these models, we can see that methods excelling at interpolation, like k-NN and Decision Trees, struggle with true extrapolation scenarios. Linear Regression provides a simple but limited global trend, while SVM produces the most realistic long-term forecasts.

## 5. Evaluation and tuning of the models

To evaluate our models, we first split the dataset into training and test sets using a randomized 80/20 split. Each model is trained on the training portion and then used to predict the values in the test set. We assess prediction quality using several complementary metrics:

- MAE (Mean Absolute Error), which measures the average absolute deviation;
- RMSE (Root Mean Squared Error), which penalizes larger errors more strongly;
- $R^2$, which indicates how much variance the model is able to explain.

This evaluation step allows us to objectively compare the performance of Linear Regression, Decision Trees, k-NN, and SVM under identical conditions.

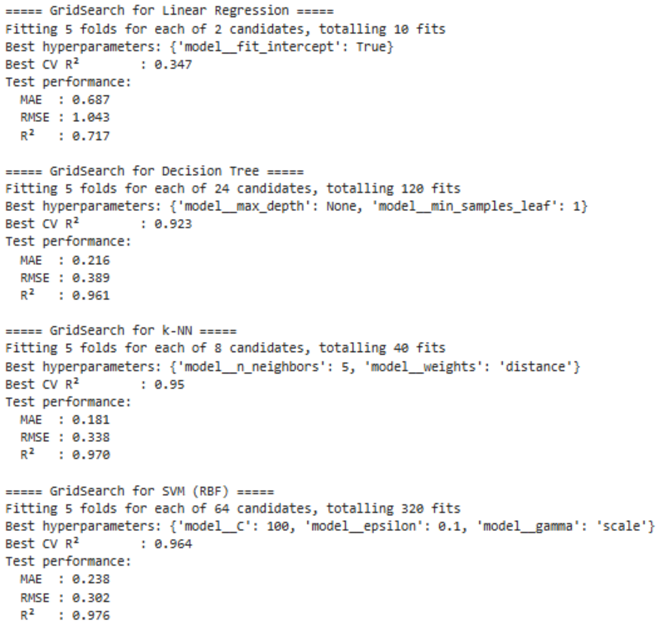


| | Model | MAE | RMSE | R$^2$ |
|---|---|---|---|---|
| 0 | Linear Regression | 0.687 | 1.043 | 0.717 |
| 1 | Decision Tree | 0.261 | 0.409 | 0.956 |
| 2 | k-NN | 0.234 | 0.458 | 0.945 |
| 3 | SVM | 0.383 | 0.691 | 0.876 |

Figure 13: Models evaluation

Although k-NN and Decision Trees achieve better metrics on a random train test split as seen above, this is because they excel at interpolation: they memorize historical points very well and can easily predict years close to those already seen. However, when looking at projections up to 2050, we are in a true extrapolation scenario, where k-NN and Decision Trees completely fail (producing constants). In contrast, the SVM, manages to extend the observed dynamics correctly and produces much more plausible curves. Thus, even if SVM is not always the best in terms of metrics, it is the model that captures long-term trends most accurately.

Next, we use `GridSearchCV` to automatically search for the optimal hyperparameters of each model. We build pipelines that include preprocessing steps such as `StandardScaler` when necessary and define parameter grids for each algorithm, for example different tree depths for the Decision Tree, different numbers of neighbors for k-NN, or various values of $C$, $\gamma$, and $\epsilon$ for the SVM. `GridSearchCV` tests every combination and selects the one that achieves the highest average $R^2$. After identifying the best configuration, we retrain the model on the full training set and evaluate it again on the test data. This systematic approach ensures that each model is used in an optimized form, making the comparison fair and strengthening the reliability of the conclusions drawn from our forecasting results.



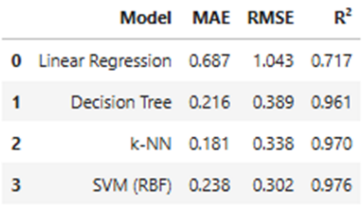
```
===== GridSearch for Linear Regression =====
Fitting 5 folds for each of 2 candidates, totalling 10 fits
Best hyperparameters: {'model__fit_intercept': True}
Best CV R²         : 0.347
Test performance:
  MAE  : 0.687
  RMSE : 1.043
  R²   : 0.717

===== GridSearch for Decision Tree =====
Fitting 5 folds for each of 24 candidates, totalling 120 fits
Best hyperparameters: {'model__max_depth': None, 'model__min_samples_leaf': 1}
Best CV R²         : 0.923
Test performance:
  MAE  : 0.216
  RMSE : 0.389
  R²   : 0.961

===== GridSearch for k-NN =====
Fitting 5 folds for each of 8 candidates, totalling 40 fits
Best hyperparameters: {'model__n_neighbors': 5, 'model__weights': 'distance'}
Best CV R²         : 0.95
Test performance:
  MAE  : 0.181
  RMSE : 0.338
  R²   : 0.970

===== GridSearch for SVM (RBF) =====
Fitting 5 folds for each of 64 candidates, totalling 320 fits
Best hyperparameters: {'model__C': 100, 'model__epsilon': 0.1, 'model__gamma': 'scale'}
Best CV R²         : 0.964
Test performance:
  MAE  : 0.238
  RMSE : 0.302
  R²   : 0.976
```

Figure 14: GridSearchCV analysis for our models



|  | Model | MAE | RMSE | $R^2$ |
|---|---|---|---|---|
| 0 | Linear Regression | 0.687 | 1.043 | 0.717 |
| 1 | Decision Tree | 0.216 | 0.389 | 0.961 |
| 2 | k-NN | 0.181 | 0.338 | 0.970 |
| 3 | SVM (RBF) | 0.238 | 0.302 | 0.976 |

Figure 15: Final model result

Hyperparameter tuning improved all non-linear models, while the linear regression remained unchanged (no meaningful parameters to optimize). The Decision Tree shows a small but stable gain in accuracy, with $R^2$ increasing from 0.956 to 0.961. The k-NN model improves more significantly thanks to optimized neighborhood size and weighting, raising $R^2$ from 0.945 to 0.970. The largest improvement comes from the SVM (RBF), where tuning $C$, $\gamma$ and $\epsilon$ reduces the error by half and boosts $R^2$ from 0.876 to 0.976, making it the best-performing model. In summary, hyperparameter tuning clearly benefits the non-linear models, with the tuned SVM achieving the highest overall performance.

For our analysis of the evolution of renewable energy share for 2050, we can say that the best model is the SVM although model metrics such as MAE, RMSE, and $R^2$ are not always the best. These metrics provide useful information about how well each algorithm fits the test split, but they do not necessarily indicate how reliable the model will be for long-term forecasting. A model can perform very well on randomly shuffled test data while still producing unrealistic or unstable projections for years far beyond the training range. This is why the final choice is not based solely on numerical performance: the shape and behavior of the projection curve also matter. In this case, the SVM model was selected because its extrapolation pattern is smoother and more consistent with the historical trend, even if other models achieved slightly better scores on the 80/20 split.

Finally, we applied the SVM model to each income group to predict future renewable energy trends. For every group, we treat the year as the input feature and the average renewable share as the target. A randomized 80/20 train test split ensures that the model sees a balanced mix of years during training. The SVR is then fitted, and the model is used to generate predictions from the earliest year in the dataset up to 2050.

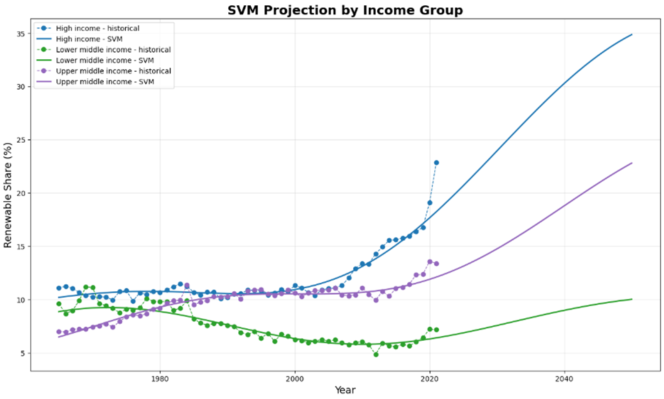


Figure 16: SVM projection by income group

The resulting graph highlights clear structural differences between income groups. High-income countries start with the highest renewable share and, according to the SVM, are projected to accelerate strongly in the coming decades. Upper-middle-income countries show a more gradual but steady increase, with the model suggesting a significant catch-up compared with their initial values. By contrast, lower-middle-income economies display a flatter trajectory: their historical averages are lower and the SVM projection remains more modest, indicating slower growth in the renewable share if current patterns persist. Overall, this analysis suggests that share of renewables is closely linked to both the incomes and the expected future speed of the energy transition.

High-income countries show the steepest projected increase, which can be explained by their greater investment capability, faster adoption of new technologies, and stronger policy commitments toward decarbonization. These countries have already started from higher renewable-energy levels and have the financial means to accelerate this transition further. Upper-middle-income economies exhibit a moderate upward trend, reflecting improving infrastructure but more limited resources compared to high-income regions. Meanwhile, lower-middle-income countries show a flatter projection: their slower growth is typically linked to structural barriers such as restricted capital, lower electrification rates, and competing development priorities. Overall, the SVM projection reinforces the idea that economic development strongly influences a country's ability to scale up renewable energy in the coming decades.

When comparing these projections with global climate-change objectives, a clear gap emerges between expected trends and what would be required to meet international targets. According to the IPCC and the IEA, the share of renewable energy in the global mix must rise sharply, reaching well above 60–70% by 2050 to keep warming below 1.5 °C. The SVM results show that high-income countries are the only group following a trajectory that approaches this pace of expansion, driven by stronger policy support, mature infrastructure, and higher investment capacity. Upper-middle-income economies make progress but remain significantly below the required acceleration, suggesting that their current efforts are insufficient to align with a 1.5 °C pathway. Lower-middle-income countries fall even further behind, with projections indicating only modest improvements that would not meaningfully contribute to global decarbonization goals. Overall, the comparison highlights a structural imbalance: without substantial international support, financial transfers, and technology sharing, the pace of renewable-energy adoption in lower-income groups will remain too slow to collectively meet global climate targets.

## 6. Conclusion

This project combined data cleaning, exploratory analysis, and several machine-learning models to investigate global renewable energy trends and produce projections to 2050. After constructing a clean dataset, we generated historical averages and trained models such as linear regression, k-NN, decision trees, and SVMs. Among these, the SVM consistently provided the most realistic long-term trajectories, successfully capturing the non-linear growth patterns visible in the data. By applying this model separately to each income group, we highlighted strong structural disparities: high-income countries show rapid and accelerating renewable-energy adoption, upper-middle-income countries progress more steadily, while lower-middle-income groups remain far behind.

These projections, however, also reveal a concerning difference with global climate objectives. Even under optimistic assumptions, the modeled growth of renewables, especially in middle-income regions, remains insufficient to meet the levels required to keep global warming below 2 °C, let alone 1.5 °C. High-income countries alone cannot compensate for the slower transition elsewhere, and the aggregated global trend implied by our models corresponds to a warming pathway above 2 °C. This highlights a fundamental challenge: achieving climate targets will require not only continued acceleration in developed economies, but also significant international support, investment, and technological diffusion to enable faster renewable expansion in emerging and developing countries.

Overall, the project demonstrates the value of combining statistical preprocessing with machine-learning models to explore energy-transition pathways, while also making clear that current trends, even when optimistically extrapolated, fall short of what is needed to stabilize the climate.