Power Plant Generation Forecasting: A Comparative Study of Linear Regression, Random Forest and KNN

Buontempi Andrea, Bursi Nicole, Sepe Raffaele

1. Introduction to the Problem

Electricity represents a backbone of modern society, and its continuous availability is a fundamental prerequisite for industrial activities, essential services, and overall quality of life. In this context, power system management depends not only on the installed capacity of power plants but, more crucially, on the ability to reliably forecast how much energy will be produced and within which time intervals.

Despite the critical importance of these forecasts, the data required to build robust models is not always immediately accessible or consistent. Information regarding power plants (such as location, nominal capacity, technology, and fuel type) is often scattered across multiple sources. Similarly, historical generation time series may be incomplete, collected using differing criteria, or available only for specific geographical areas and timeframes.

Our research addresses this specific challenge: developing a forecasting approach for power plants by using structural asset data and available historical records to reconstruct and anticipate the evolution of generation over time. The objective is to build a model capable of capturing both temporal dynamics (trends, seasonality, cyclicity, and shocks) and the heterogeneity between different plants and countries, ultimately producing estimates that are valuable for energy analysis and decision support.

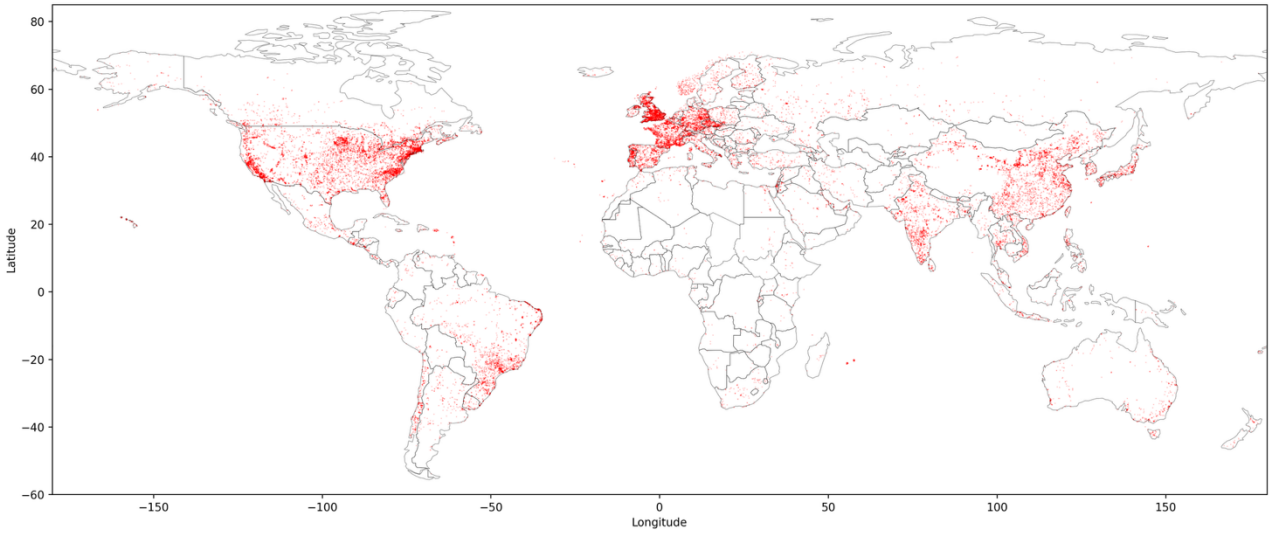
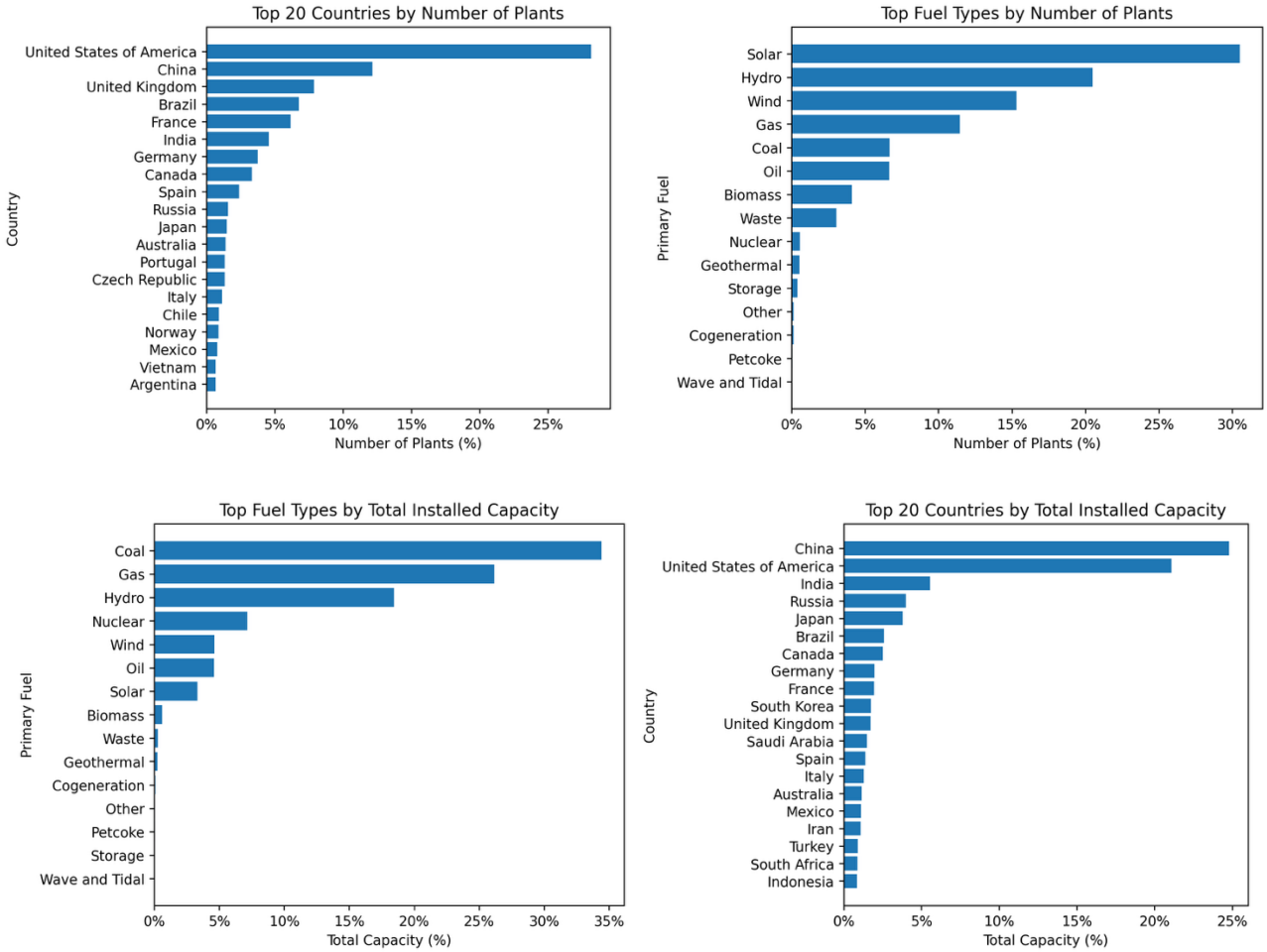


Figure 1: Geographic Distribution of Power Plants

1.1 Input File

The analyses conducted for this report are based on a single .csv file entitled *global_power_plant_database.csv*. This is an open-source dataset developed by the *World Resources Institute (WRI)* and its partners. It integrates hundreds of public sources and contains technical specifications for power plants, such as installed capacity, location, and fuel type, while also including power generation data (GWh) where publicly available. The *Global Power Plant Database (GPPD)* comprises information on approximately 35,000 power plants across 167 countries, representing roughly 72% of the world's total installed capacity.

The following graphs illustrate the statistics by country and plant type, based on the data provided in the .csv file:



2. Proposed Approach

To achieve an optimal resolution of the problem, we adopted the following approach: a three-year analysis (2017, 2018, and 2019) was conducted using a comparative framework. For each forecast year, we compared three models, each characterized by different assumptions and strengths:

1. Linear Regression
2. Random Forest
3. K-Nearest Neighbors (KNN)

Specifically, these three models are central to performing forecasts across three distinct levels of granularity:

- The first approach focuses on individual power plants.
- The second examines all power plants within a specific country, aggregating data by nation.
- Finally, the third forecasts generation by aggregating all plants based on their fuel type.

The results were subsequently evaluated and compared based on two performance metrics, the Mean Squared Error (MSE) and the Mean Absolute Error (MAE), in order to determine the most effective model for each specific analytical level.

2.1 Description of the State of Art

The literature on energy forecasting and power systems frequently proposes a comparison between Linear Regression, K-Nearest Neighbors, and Random Forest as benchmark models to evaluate approaches with varying degrees of interpretability and different capabilities for capturing non-linear dynamics. For instance, in predicting the output of conventional power plants (such as combined-cycle plants), these models are compared to determine when a linear model is sufficient and when more flexible methods are required.
In short-term load forecasting, KNN variants prove to be effective when demand exhibits recurring patterns,

as the forecast leverages the "similar days" concept; finally, for renewable energy generation (e.g. wind power prediction), Random Forest is widely adopted for its robustness in capturing non-linearities and interactions between variables.
A prevalent approach in the state of the art consists of aggregating data from multiple years into a single dataset and selecting a single "best" global model based on the average performance across the entire period. However, this approach implicitly assumes that the relationship between features and targets is substantially stable over time. Consequently, the selection based on an average may mask differences between sub-periods (e.g., years with different operating conditions or contexts), resulting in a model that is "good on average" but not necessarily optimal for specific time intervals.
Another line of research, represented by the papers used for this project, focuses on generation forecasting for power plants by addressing heterogeneity across plant types. Instead of applying the same model to all the plants, different models are adopted depending on the plant's primary fuel, assuming that each class has its own dynamics and distinct input-output relationships.
This highlights how the state of the art handles diversity either through a single global model or through models specialized by plant type. Our contribution falls within the first category, proposing a comparison between multiple algorithm: instead of assuming stationarity and choosing a single overall model, the idea is to evaluate the performance without flattening temporal variability. This allows us to identify when a specific algorithm is more suitable in certain years or sub-periods compared to others.

2.2 Data Processing

In order to perform an accurate and effective forecast for the three years under consideration, we started with the aforementioned *.csv* file and conducted a selection of the most relevant information. This process allowed us to build the database that was then used as input for the models selected for the analysis.

Specifically, we decided to remove several columns that were not necessary for the study:

- *other_1-3*;
- *owner;*
- *source;*
- *url;*
- *geolocation_source;*
- *wepp_id;*
- *year_of_capacity_data*;
- *generation_data_source*;
- *estimated_generation_gwh_2013-2017*;
- *estimated_generation_note_2013-2017*.

This resulted in a more streamlined initial file, providing an excellent starting point for the construction of the final database.

2.3 Database Creation

The database was implemented by using the PostgreSQL RDBMS, managed through the pgAdmin 4 graphical interface.
The population phase has occurred by importing the source dataset, *global_power_plant_database.csv*, whose data has been structured inside a database named *PLANTS*. The obtained relational schema is organized into three main tables: country, plant, and energy.

The country table includes the following columns:

- *id_country* (text): 3-character country code corresponding to the ISO 3166-1 alpha-3 specification.
- *country_name* (text): extended form of the country designation.

The plant table includes the following columns:

- *id_plant* (text): a 10 or 12-character identifier for the power plants.
- *id_country* (text): 3-character country code corresponding to the ISO 3166-1 alpha-3 specification.
- *commissioning_year* (numeric): the year the plant entered operation, capacity-weighted where data are available.

- *primary_fuel* (text): the energy source used for primary electricity generation or export.
- *capacity_mw* (numeric): electrical generating capacity in megawatts.
- *latitude* (numeric): geolocation in decimal degrees; WGS84 (EPSG:4326).
- *longitude* (numeric): geolocation in decimal degrees; WGS84 (EPSG:4326).

The energy table includes the following columns:

- *id_plant* (text): a 10 or 12-character identifier for the power plants.
- *generation_2013_gwh*, ..., *generation_2019_gwh* (numeric): reported electricity generation in gigawatt/hours for the years ranging from 2013 to 2019.

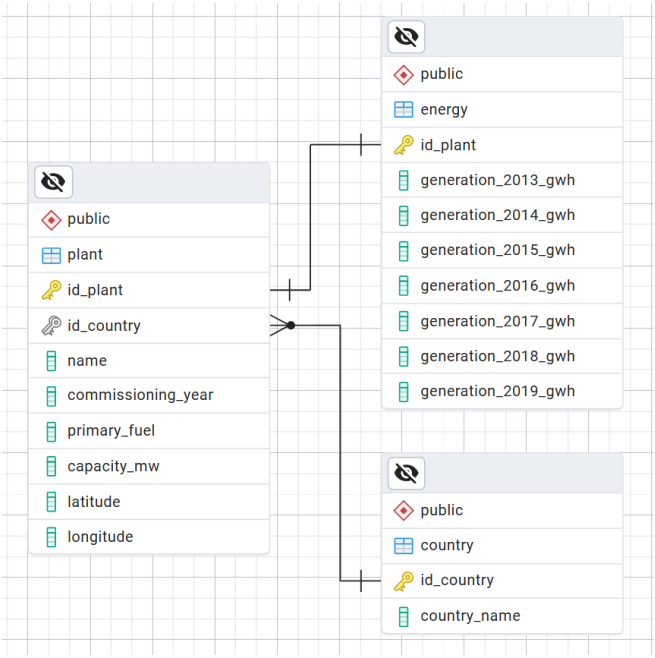


Figure 6: Logic Schema

2.4 Data Pre-Processing prior to the Forecasting Phase

The pre-processing phase was designed and differentiated according to the specific type of forecasting being addressed. We implemented three distinct methods for data cleaning and reorganization, depending on whether the forecast targeted:

1. Individual Power Plants
2. Power Plants by Country
3. Power Plants by Fuel Type

2.4.1 Data Processing at the Individual Plant Level

We implemented the following changes:

- We transformed the variables *country_name* and *primary_fuel* into categorical variables, each identified by a unique code.
- The power plant capacity was converted into gigawatts (GW).
- As regards the training data, we restricted our dataset to records with no missing values across the selected features. This methodology was justified by the fact that, even after filtering, a substantial sample of over 6,000 records was preserved for the analysis.

2.4.2 Data Processing by Country and Technology Aggregation

We implemented the following modifications:

- The power plant capacity was converted into gigawatts (GW).
- For generation columns with a total sum of zero, we imputed estimated values. These were obtained by multiplying the national capacity by the average annual full-load hours.
- Only records with complete observations for all selected features were included in the training dataset.

3. Results

All dataframes used in this study were created by directly connecting the database to Visual Studio Code. The models implemented for the forecasting analysis are: Linear Regression, Random Forest, and KNN.

3.1 Forecasting Model used for Each Plant

3.1.1 Correlation Analysis and Feature Selection

In the preliminary phase, a correlation matrix was analyzed to identify the most significant features according to the target variable. The objective was to select a subset of predictors demonstrating a strong linear dependence with the target for the reference years (2017, 2018, and 2019), thereby reducing dimensionality and data noise.

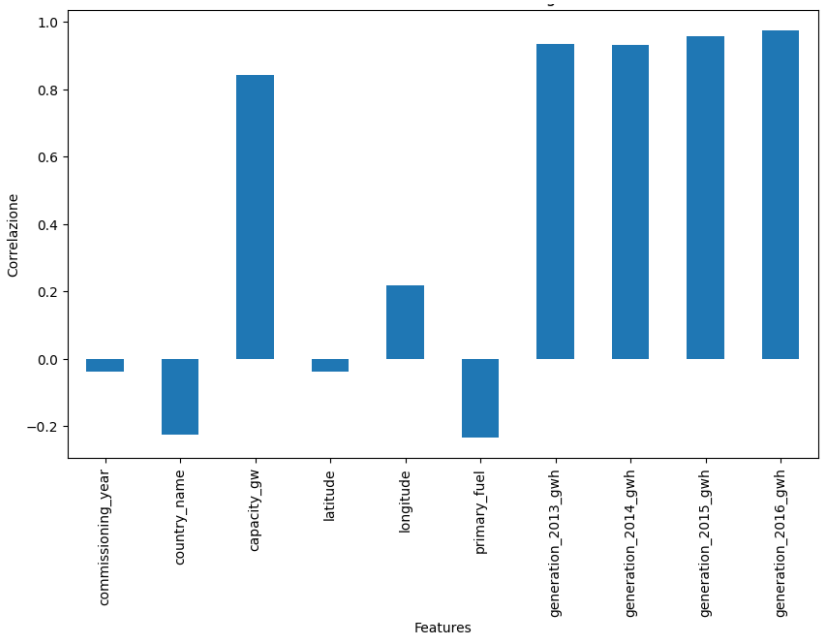


Figure 7: Correlation Between Features and *Target1*

The initial models were exclusively trained on this set of high-correlation features. It is important to note a specific detail regarding the K-Nearest Neighbors algorithm: contrary to standard practice, which typically requires data normalization or standardization for distance-based algorithms, our empirical tests showed a degradation in performance when such pre-processing was applied. Consequently, for the KNN model, we opted to use the data in their original scale.

The results of this initial experimental phase are shown in the following figure:

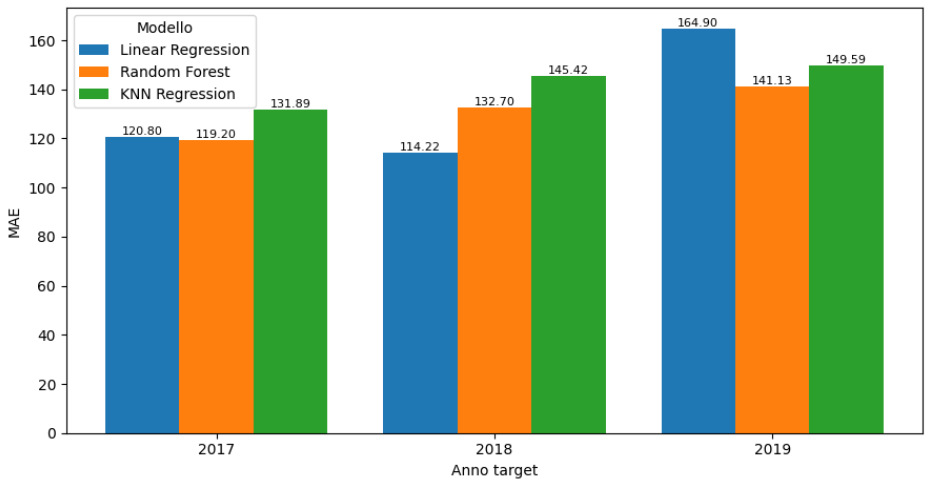


Figure 8: MAE Comparison: Model vs. Target (2017-2019)

3.1.2 Baseline Definition and Comparative Evaluation

To validate the effectiveness of the implemented Machine Learning models and determine the acceptability of the results, a deterministic baseline was introduced for comparison.

The baseline was obtained by calculating the arithmetic mean of the values observed in the years preceding the target forecast year. Formally, the prediction for a given year (e.g., 2017) was calculated as the average of the values from 2013 to 2016.

Comparing the error metrics of the predictive models against those of the baseline allows a quantification of the actual predictive gain provided by the algorithms.

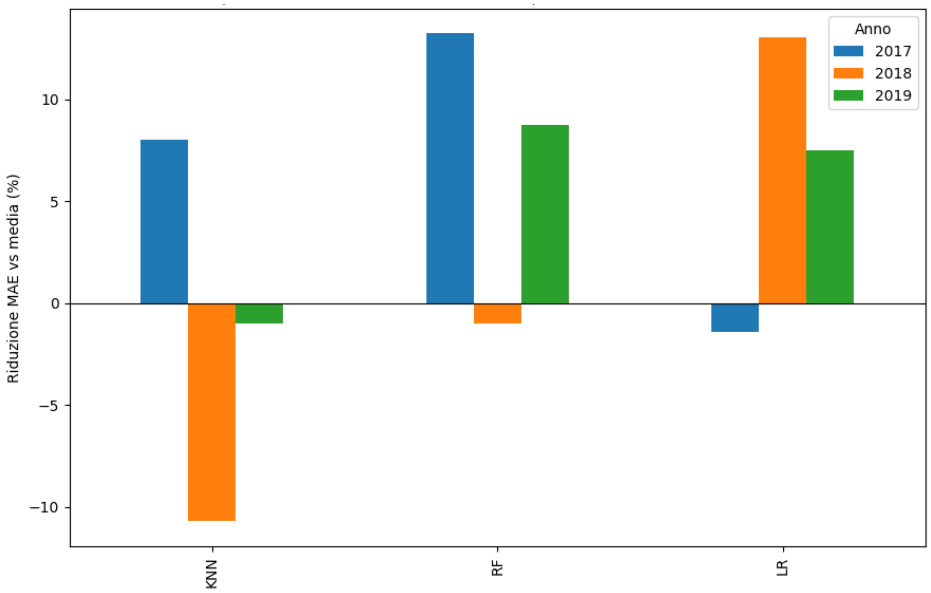


Figure 9: % Reduction in MAE of Models Relative to the Mean MAE (2017-2019)

3.1.3 Hyperparameter Tuning

Subsequently, the focus shifted on optimizing the Random Forest and KNN models to enhance their predictive performance. An automated search procedure (such as Grid Search or an iterative process) was implemented to identify the combination of hyperparameters that minimizes the objective function.

Consequently, the optimal settings for each algorithm were determined based on the specific characteristics of the dataset.

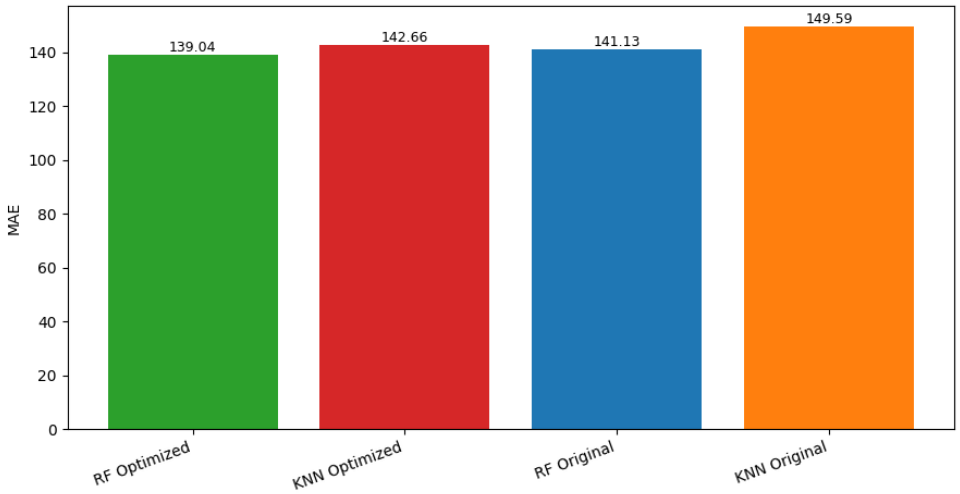


Figure 10: MAE Comparison: Original vs. Optimized RF/KNN (Target 2019)

3.1.4 Impact Analysis of All Available Features

To test the hypothesis that variables with low apparent correlation might still contain valuable information for predictions (such as non-linear relationships), an experiment was conducted by extending the training set to all available variables, including those with correlation coefficients near zero.

Training the models on this expanded feature space produced the results shown in the graph below, enabling us to assess whether greater informational complexity resulted in a tangible improvement in performance metrics.

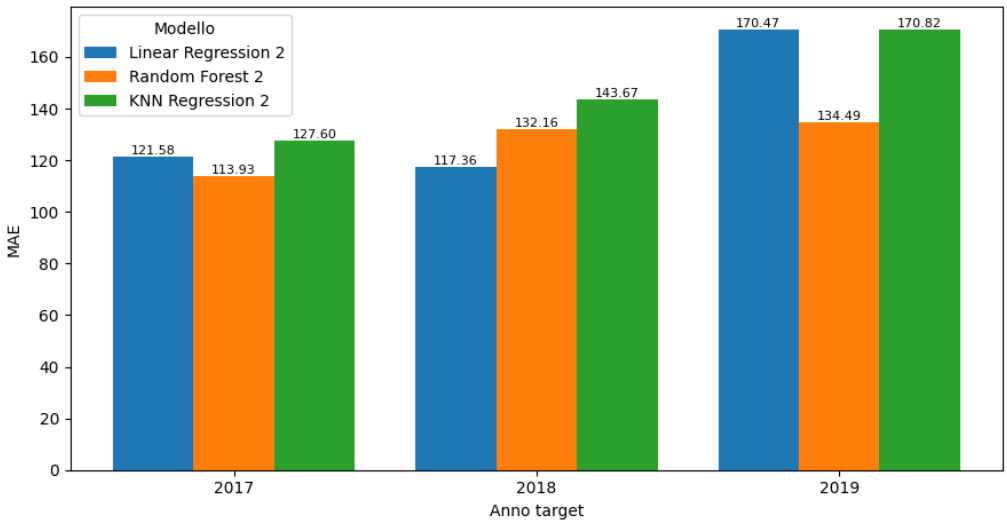


Figure 11: MAE Comparison by Model and Target (2017-2019)

3.1.5 Experimentation with Multiple Temporal Variables

Finally, an approach integrating a temporal component into the dataset was tested. The objective was to determine whether using target values from previous years (2017 and 2018) as explanatory variables (features) could improve the prediction for the 2019 target year.

Two training scenarios for the 2019 forecast were compared:

1. Reduced Dataset: high-correlation features combined with historical targets (2017, 2018).
2. Full Dataset: all available features combined with historical targets (2017, 2018).

This configuration aims to leverage the temporal autocorrelation of the data, providing the model with direct information regarding the recent trends of the phenomenon.

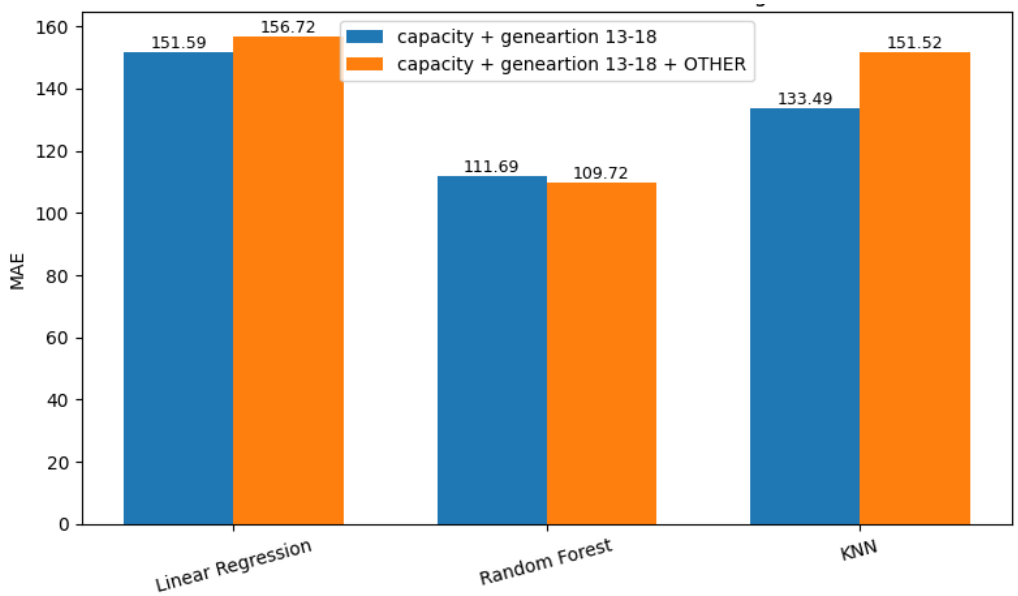


Figure 12: MAE Comparison: Comparing Different Feature Sets (Target 2019)

3.2 Forecasting Models by Country Aggregation

In this phase of the study, the dataset was reorganized by grouping records by country (geographical aggregation). In this scenario, the predictive models were trained using a subset of features highly correlated with the target. This input set comprises installed capacity (*capacity*) and historical generation records for the 2013–2016 gap period (*gen13, gen14, gen15, gen16*).

Performance evaluation was conducted using Mean Absolute Error (MAE) as the primary metric. It should be noted that no further exploratory analyses or optimization procedures (such as hyperparameter tuning or comparison with different baselines) were repeated; the applicable methodologies and related considerations would have been redundant, given the detailed discussion already provided in *Section 3.1*.

It is important to note that there is a potential risk of overfitting for the Linear Regression model in this context.

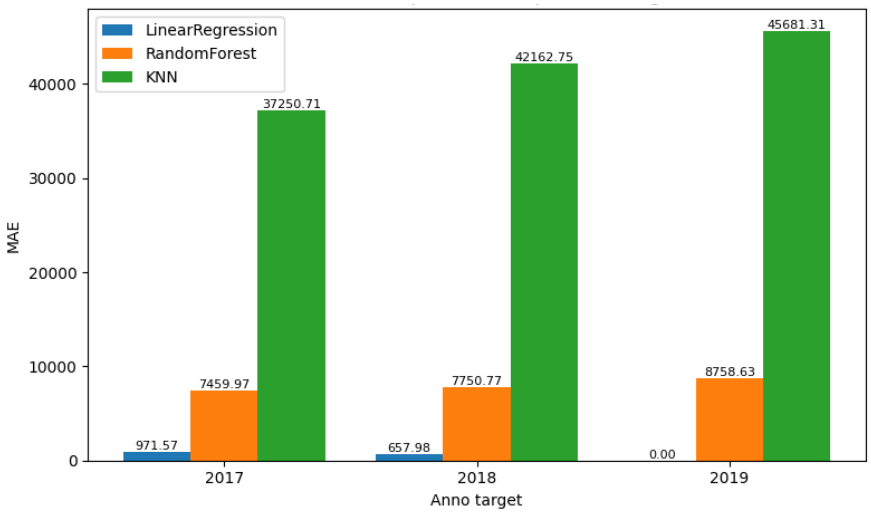


Figure 13: MAE Comparison by Model (Per Target Year)

3.3 Forecasting Models by Fuel Type

Following the same approach used for the geographical distribution, an aggregate analysis was conducted by fuel type. The dataset was grouped according to the primary energy source. For the training phase, the feature selection was limited to variables demonstrating a high correlation with the target (specifically, installed capacity and historical generation values: *gen13, gen14, gen15, and gen16*).

The model performance was evaluated using the Mean Absolute Error (MAE) metric. In this section, no further optimization tests or comparisons with complex baselines were performed, as the methodological considerations and validation procedures remain consistent with the detailed explanation provided in *Section 3.1*, making further repetition unnecessary.

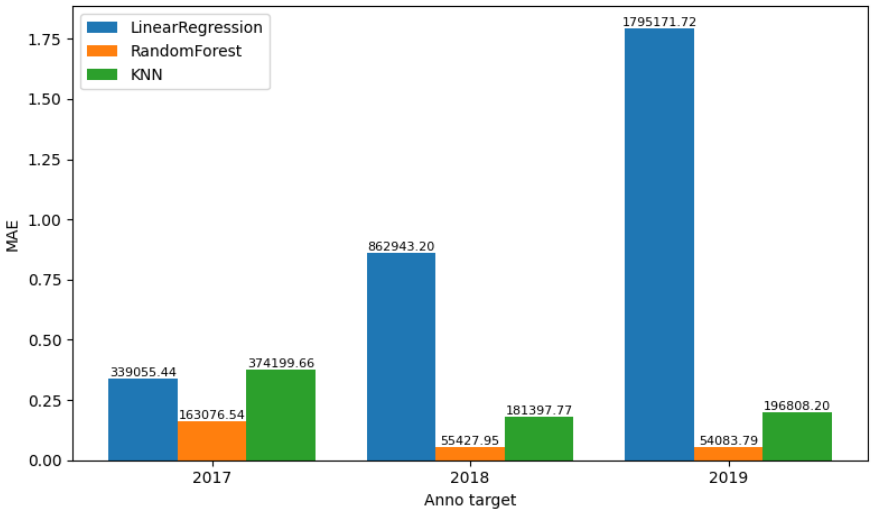


Figure 14: MAE Comparison by Model (Per Target Year)

4. Conclusions

The findings of this study lead to the following conclusions:

- The KNN model proved to be the least effective model compared to the baseline, whereas both Random Forest and Linear Regression ones demonstrated satisfactory performances.
- The year 2017 yielded the most accurate predictions. This result was expected, as predictive precision typically decreases as the forecast horizon moves further away from the training set's time window.
- Including features with low correlation slightly degraded the models' performance; the only exception was the Random Forest model, which showed an improvement, likely due to its ability to handle complex, non-linear relationships.
- When predicting 2019 by incorporating *gen17* and *gen18* as additional features, all models showed a lower error rate. This improvement is attributed to the inclusion of two extra variables that are highly correlated with the target.
- Linear Regression was the most effective model for the country-level aggregation, while Random Forest performed better for the fuel-type aggregation. This discrepancy may be attributable to the difference in sample sizes: the country-level grouping provided 167 records, whereas the fuel-type grouping consisted of only 15 records.