Final Project

Data Science Applied to Electrical Energy Systems

ANALYSING SPANISH ELECTRICITY MARKET USING MACHINE LEARNING

REPORT

Authors: Alex Blanco, Guillem Mañé, Alex Vea, Ivan Moreno and Gvidas Andzevicius

Director: Sara Barja Martínez

Group: G12 - 2

ETSEIB

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

UPC

CONTENTS

List of figures

1. INTRODUCTION

The increasing complexity of electricity markets, driven by the integration of renewable energy sources, fluctuating fuel prices, and changing weather conditions, highlights the need for accurate data analysis and prediction. Effective forecasting and pattern identification are crucial for market participants to optimize operations and make informed decisions. This project leverages advanced machine learning techniques to address these challenges in the Spanish electricity market using historical data from the ESIOS platform provided by Red Eléctrica de España (REE).

The project focuses on three main tasks: supervised learning to predict day-ahead electricity prices, unsupervised learning to cluster daily price profiles and identify patterns, and dimensionality reduction using Principal Component Analysis (PCA). These methodologies provide a comprehensive approach to understanding and forecasting market dynamics.

Initial exploratory data analysis (EDA) was conducted to ensure the dataset was clean, visualized, and ready for modelling. The supervised learning model was rigorously evaluated using metrics such as RMSE, MAE, and R^2, while clustering methods like K-means and hierarchical clustering revealed significant patterns in price behaviours. PCA was applied to reduce data complexity and assess its impact on model performance, enhancing computational efficiency.

This report synthesizes these methodologies to offer a structured framework for applying machine learning to electricity market data. The findings contribute to the growing discourse on utilizing data science to improve energy system operations and support decision-making in an evolving market landscape.

2. EDA

The Exploratory Data Analysis (EDA) is a critical first step in any data-driven project. It involves systematically examining the dataset to understand its structure, characteristics, and potential issues. The main goals of EDA are to uncover patterns, spot anomalies, test hypotheses, and verify assumptions through summary statistics and graphical visualizations. This process helps identify key variables, relationships, and trends that will guide feature selection and model development. In the context of this project, EDA focuses on understanding the factors influencing electricity prices, such as historical price data, renewable generation, and external variables like natural gas prices, ensuring the data is prepared and reliable for machine learning applications.

As a preliminary step in our analysis, we can visualise the hourly electricity price (target variable) over the period from October 2022 to September 2024 in Figure 1. The graph reveals significant variability in electricity prices, with noticeable fluctuations across the entire timeline. Peaks in price are observed intermittently, suggesting periods of high demand or reduced supply. Conversely, there are troughs indicating moments of lower prices, likely driven by increased renewable generation or lower demand.

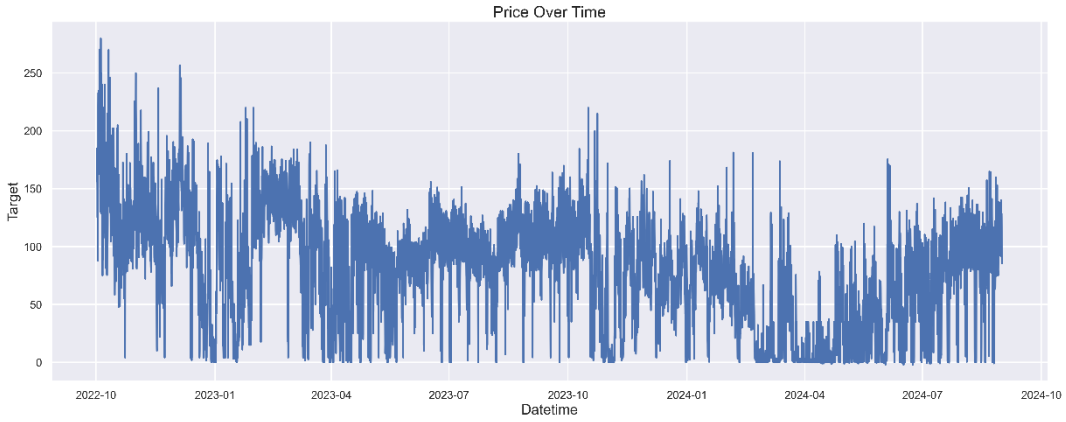

Figure 1. Electricity price over time

Seasonal patterns may also be inferred from the data, with potential price dips during certain months, likely correlating with seasonal factors such as weather conditions or changes in energy consumption habits. This visualization highlights the dynamic nature of the electricity market and serves as a foundation for further analysis to identify underlying drivers and predict future trends.

2.1. Boxplots

The boxplots provide a detailed breakdown of electricity price distributions across different temporal dimensions, offering additional insights into the variability and patterns of the target variable.

The year distribution boxplot reveals a decline in median electricity prices from 2022 to 2024, alongside a reduction in the interquartile range. This trend likely reflects evolving market conditions, such as increased renewable energy penetration or policy changes impacting price dynamics. Also, we can note a significant seasonal variability in month distribution boxplot. Months like January and February show higher price medians and broader ranges, while summer months, particularly July and August, exhibit reduced variability and generally lower medians. This reflects seasonal trends in demand and renewable generation availability.

The week distribution demonstrates fluctuations in price variability across the year. Early weeks display wider interquartile ranges, indicating more volatile prices, while mid-year weeks tend to exhibit narrower ranges. The end-of-year weeks see an increase in both median prices and variability, potentially driven by higher winter demand. Finally, the hour distribution boxplot underscores the daily cycle of electricity prices. Prices tend to peak in the late afternoon and evening hours, coinciding with high demand periods, while early morning hours exhibit lower medians and reduced variability. These patterns reflect typical daily demand cycles in the electricity market.

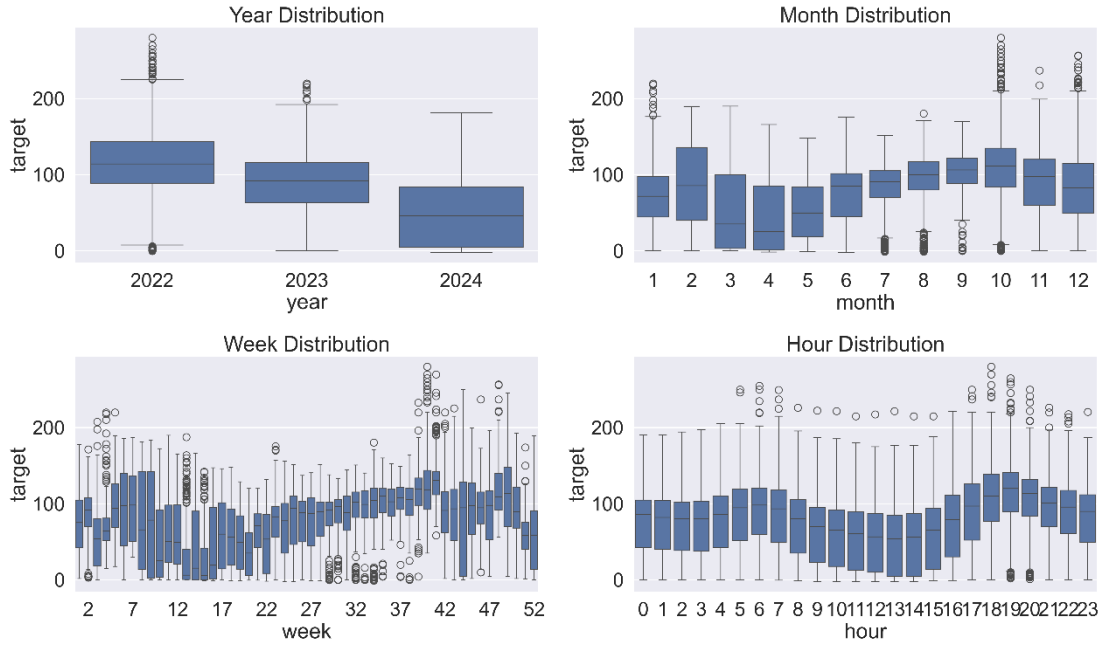

Figure 2. Year, month, week and hour distribution boxplots

2.2. Density plots

The density plot of Figure 3 illustrates the distribution of electricity prices by day of the week and year. Weekdays show wider spreads and slightly higher peaks, reflecting greater price variability due to higher demand. In contrast, weekends exhibit narrower distributions and lower peaks, indicating more stable prices.

Yearly differences reveal that 2024 experienced greater variability compared to 2022 and 2022, which align more closely with the overall distribution, suggesting price stabilization. This plot highlights the combined influence of daily and yearly temporal factors on electricity price trends, essential for refining forecasting models.

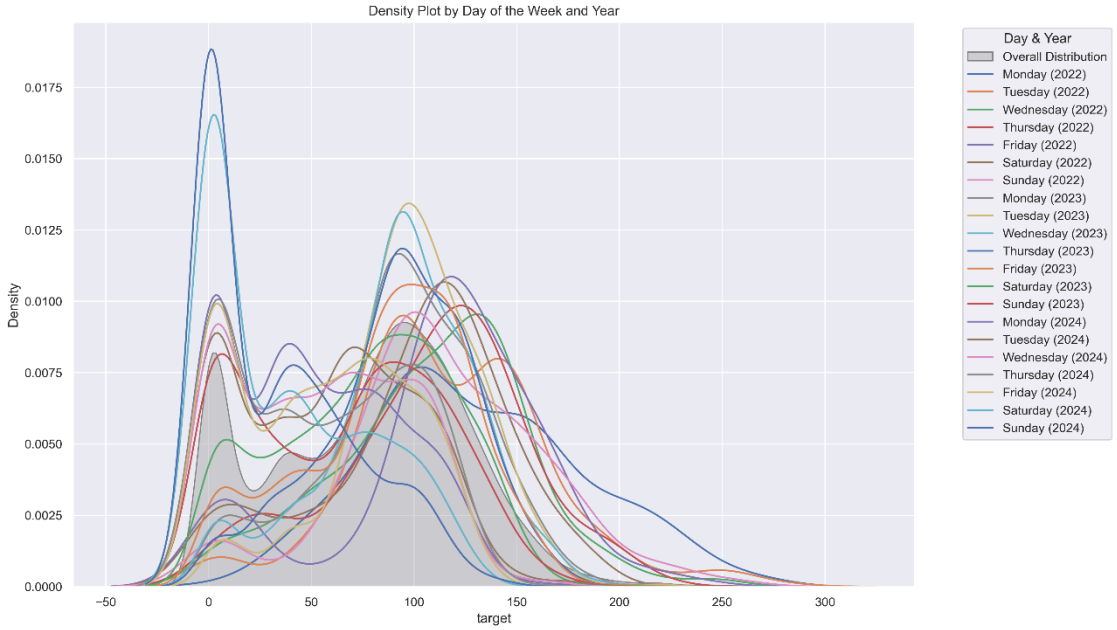

Figure 3. Density plot by day of the week and year

In contrast, Figure 4 shows electricity price distributions segmented by month and year, highlighting seasonal trends and variability across months. Notably, data for 2024 is not included due to its variability.

Some months exhibit higher densities at elevated price ranges, likely reflecting increased winter demand and lower renewable generation. In contrast, summer months display narrower distributions and lower peaks, indicating lower prices driven by higher renewable generation and reduced demand.

Figure 4. Density plot by month and year

Finally, density plot of Figure 5 represents electricity price distributions segmented by the day of the week. The overall distribution serves as a baseline, while the individual curves highlight differences in price behaviour across days.

This visualization reinforces the role of daily demand cycles in shaping electricity prices and highlights the predictable reduction in prices during weekends, essential for understanding and modelling market dynamics.

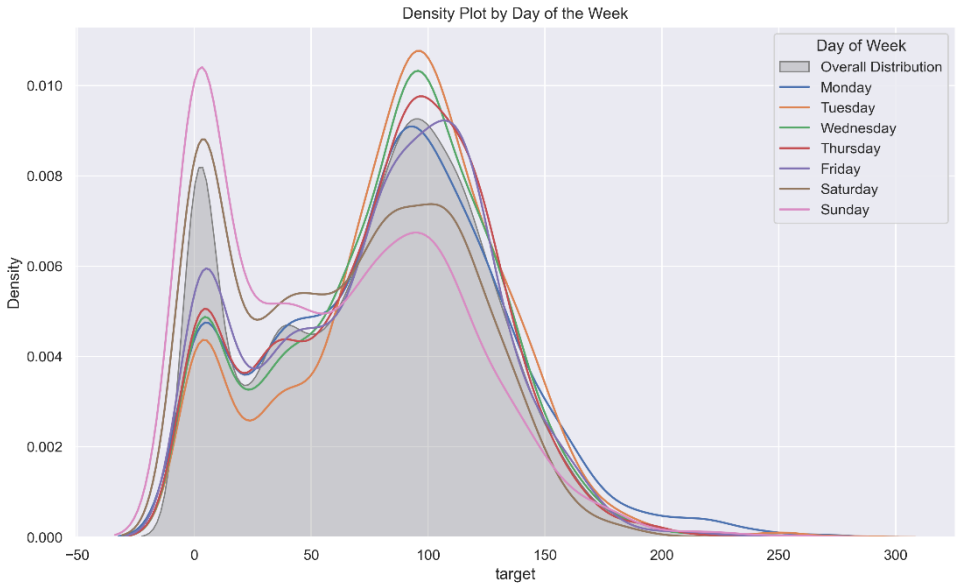

Figure 5. Density plot by day of the week

2.3. Correlation Matrix

A correlation matrix is a visualization tool used to display the strength and direction of the linear relationship between multiple variables in a dataset. Each cell in the matrix represents the correlation coefficient between two variables, with values ranging from -1 to 1. A correlation coefficient of 1 indicates a perfect positive correlation, meaning both variables increase together, while -1 indicates a perfect negative correlation, where one variable increases as the other decreases. A value close to 0 suggests little to no linear relationship. In the matrix, these relationships are often color-coded, with warmer tones (reds) indicating strong positive correlations and cooler tones (blues) representing strong negative correlations.

In the context of our dataset, several notable relationships emerge in Figure 6. The target variable, representing electricity price, shows a strong positive correlation with the previous day's target *target-1*. However, as the lag increases, the correlation progressively weakens, reflecting a diminishing influence of older price data on current values. The variable *targetmean_7day*, which represents the average electricity price over the last seven days, exhibits a strong positive relationship with the target, highlighting the importance of weekly trends in price forecasting. Similarly, the maximum and minimum values of previous targets also show significant correlations, suggesting these summary statistics are critical indicators for price prediction.

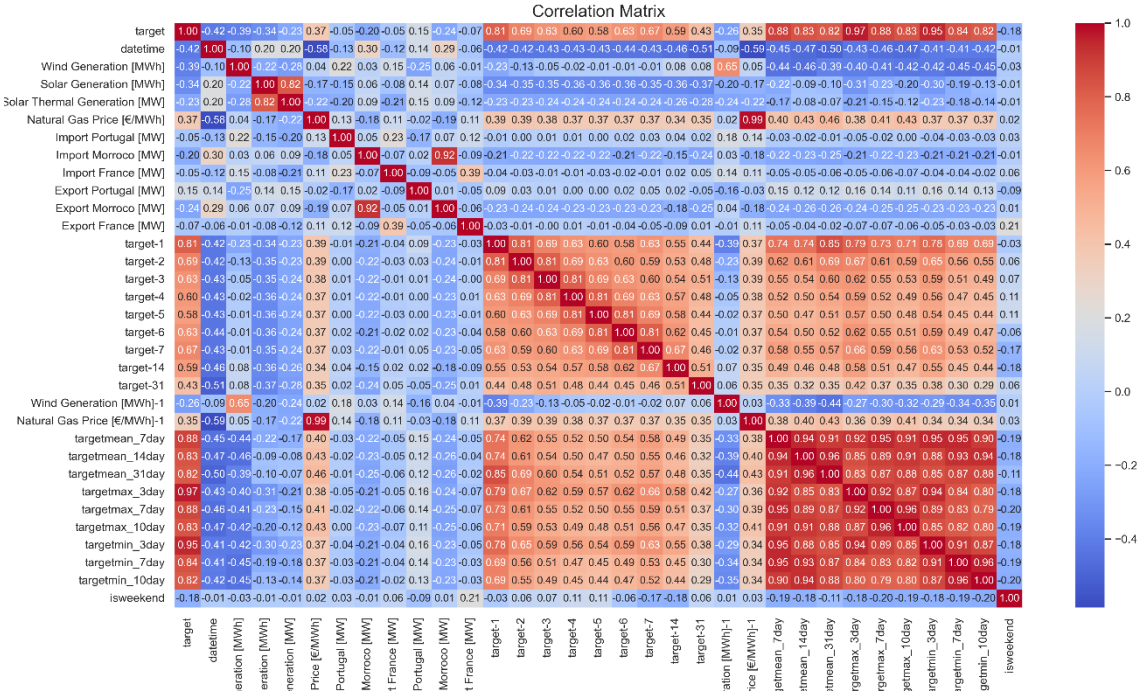

Figure 6. Correlation Matrix

The matrix also reveals an interesting dependency between electricity prices and external factors. Natural gas prices exhibit a notable positive correlation with the target, emphasizing their influence on electricity market dynamics, likely due to their role in generation costs. In contrast, renewable energy generation, particularly from wind and solar, shows a negative correlation with electricity prices. This relationship indicates that increased renewable generation tends to lower electricity prices, likely due to reduced reliance on more expensive, conventional generation methods.

Figure 7 and Figure 8 illustrate the strongest positive and negative correlations of various variables with the target variable, which represents the electricity price. They offer an insightful view into the relationships between the target and the other variables in the dataset. Positive correlations are dominated by lagged price metrics, emphasizing the strong temporal dependency of electricity prices on recent values. Aggregated metrics also show high correlations, reflecting the importance of weekly trends. Additionally, external factors like the natural gas price and cross-border flows underline the influence of fuel costs and international electricity exchanges on market dynamics.

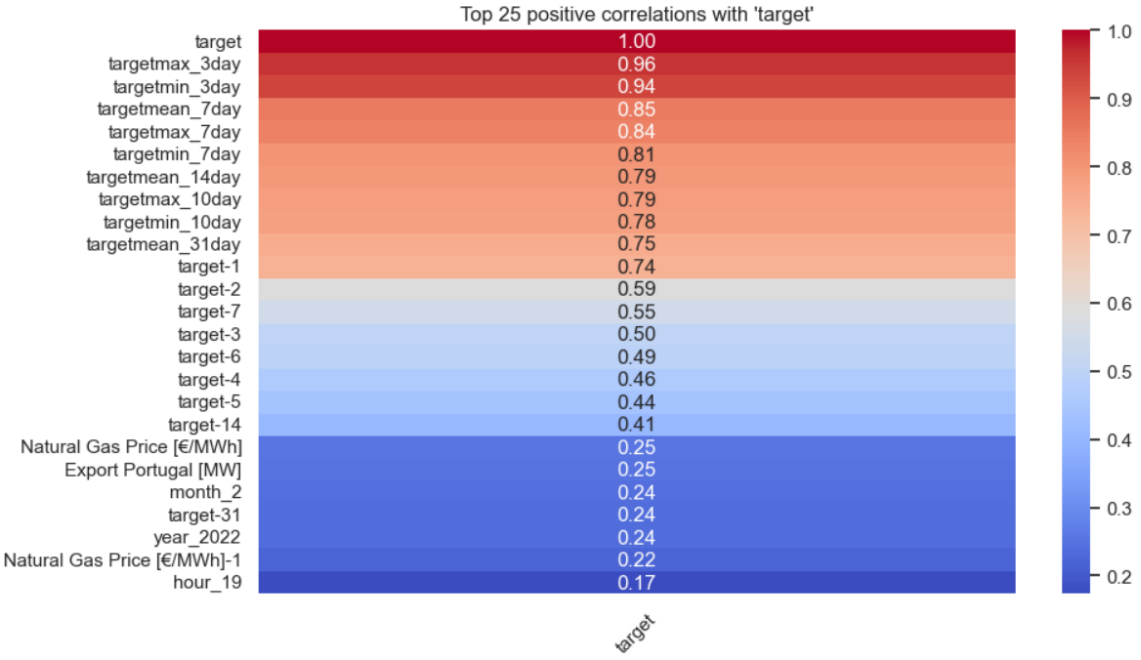

Figure 7. Top 25 positive correlations with the target

Negative correlations primarily highlight the impact of renewable energy and temporal patterns. Wind and solar generation exhibit strong inverse relationships with electricity prices, as higher renewable output reduces market costs. Temporal features confirm lower prices during weekends, while specific time periods and cross-border flows further demonstrate the interplay between energy demand, supply, and international trading. These insights are crucial for understanding and modelling electricity price behaviour.

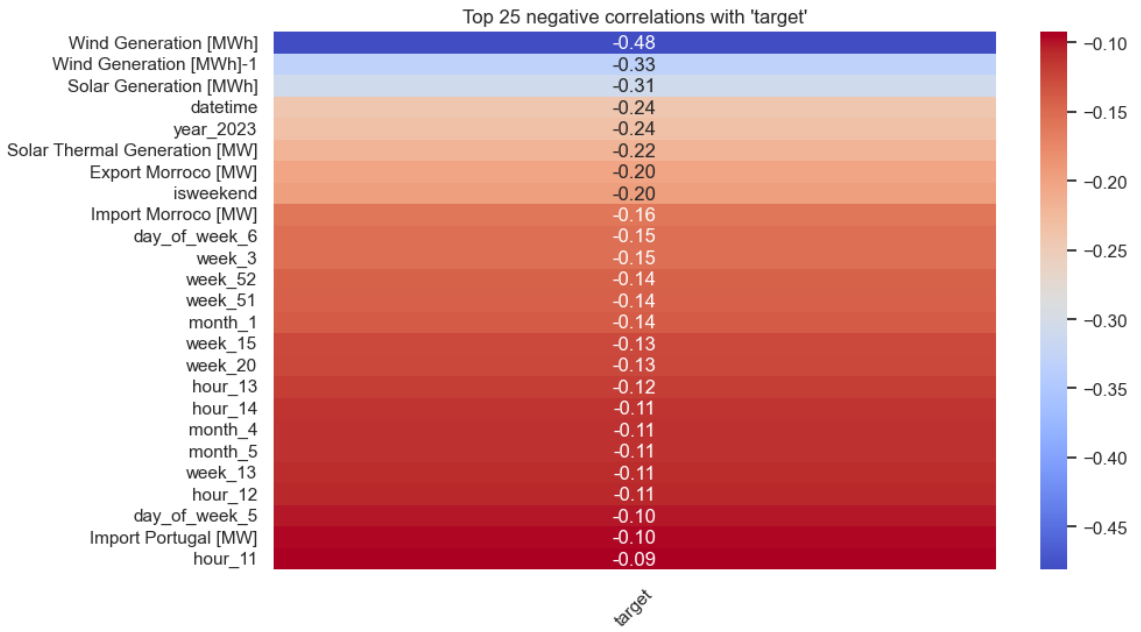

Figure 8. Top 25 negative correlations with the target

3. SUPERVISED LEARNING

Regression in supervised learning focuses on predicting a continuous output variable based on input features. Unlike unsupervised learning, where the goal is to find patterns or clusters in unlabelled data, regression uses labelled data where the output variable is known.

For example, in a machine learning model designed to predict a temperature on a room, the temperature is the continuous output being estimated.

The objective in regression is to identify the relationship between input features and the target variable, allowing the model to make accurate predictions for new data. This process not only enables predictions but also gives information into how different features influence the target. In cases where initial patterns are unclear, preprocessing steps like feature engineering or even unsupervised clustering can enhance the understanding of the data, which can later feed into a regression model to improve its performance.

In this study, the objective is to predict the SPOT price for Spain for all hours of the following day. To achieve this, historical data will be obtained from the ESIOS portal, which provides reliable and comprehensive information on electricity market prices.

The practice involves training multiple machine learning models on the historical dataset to evaluate their performance. By comparing these models, the one that achieves the best predictive accuracy with this specific dataset will be selected for forecasting.

3.1. Methodology

In this section, we describe the methodology used to apply a supervised learning approach for solving a single-value prediction problem. The process begins with the selection of all relevant features from the dataset, including those generated through feature engineering. During this step, we also filter out data from specific years identified as having excessive noise, ensuring a cleaner dataset. For categorical variables, we apply one-hot encoding to transform them into a format suitable for machine learning models.

Next, the data is split into training, validation, and test sets. We train multiple machine learning algorithms to identify the model that achieves the best performance, as measured by the R^2 metric. This is accomplished using time series split and cross-validation techniques to ensure robust evaluation. Several iterations are performed, adjusting parameters such as the test size and validation size, to optimize the training process. The model with the best performance and the most suitable hyperparameters is then selected.

Finally, the chosen model is trained using the training data and evaluated on the test set to assess its predictive accuracy.

3.2. Results

3.2.1. First iterations performed

The results from the initial iterations were not particularly satisfactory. This is evidenced by average R^2 values close to 0, indicating that the model performed no better than simply predicting the mean of the actual values.

In some cases, negative R^2 values were observed, highlighting that the model's predictive power was even worse than a simple mean prediction.

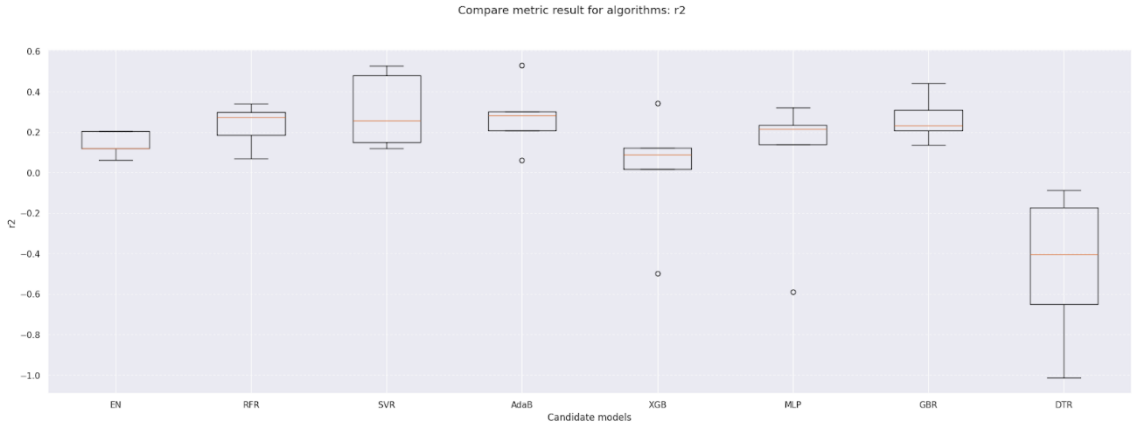

Figure 9. R^2 boxplots

Additionally, RMSE (negative root mean squared error) metrics were calculated, being equally unsatisfactory, with values ranging between -30 and -40. While the specific magnitude of these values may not seem directly meaningful, it is important to note that RMSE shares the same unit as the target variable. In this case, the unit is €/MWh, which can represent significant discrepancies in the context of energy trading environments.

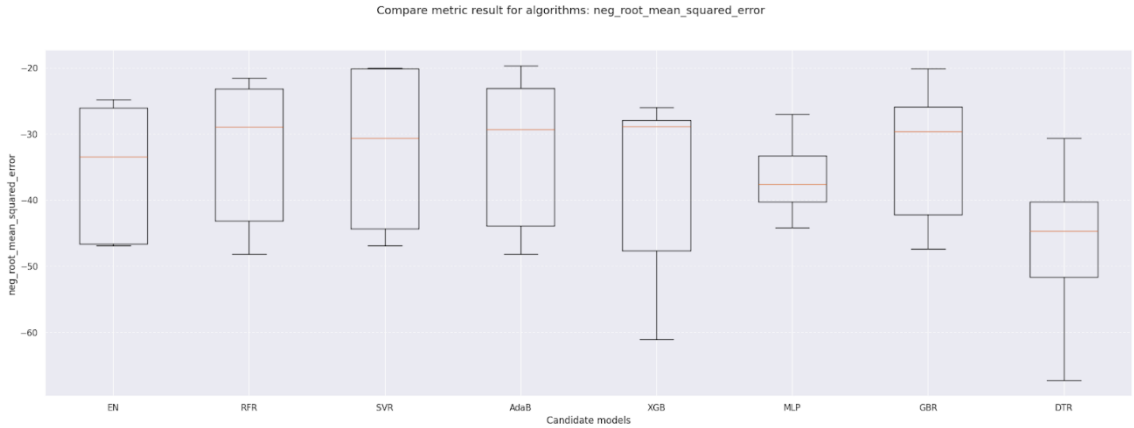

Figure 10. RMSE boxplots

3.2.2. Final model

To arrive at the final model, and considering the limitations observed in previous results, we hypothesized that the input data was not sufficiently significant for the model to effectively identify patterns. Additionally, we suspected that there might be an issue with the distributional alignment between the training and test datasets.

To investigate this, we evaluated the data distributions for the variables with the highest correlation between the test and train datasets. This analysis revealed a significant distributional issue with the dataset used.

For the final model, we ensured much greater alignment in the distributions between the training and test datasets. This improvement was achieved by modifying the number of folds used during the model's training process and by incorporating additional variables with very high correlations into the input dataset. These adjustments allowed the model to leverage more meaningful data and improve its predictive performance.

The two images presented bellow show the density plots for one feature of the training dataset, separated into test and train groups.

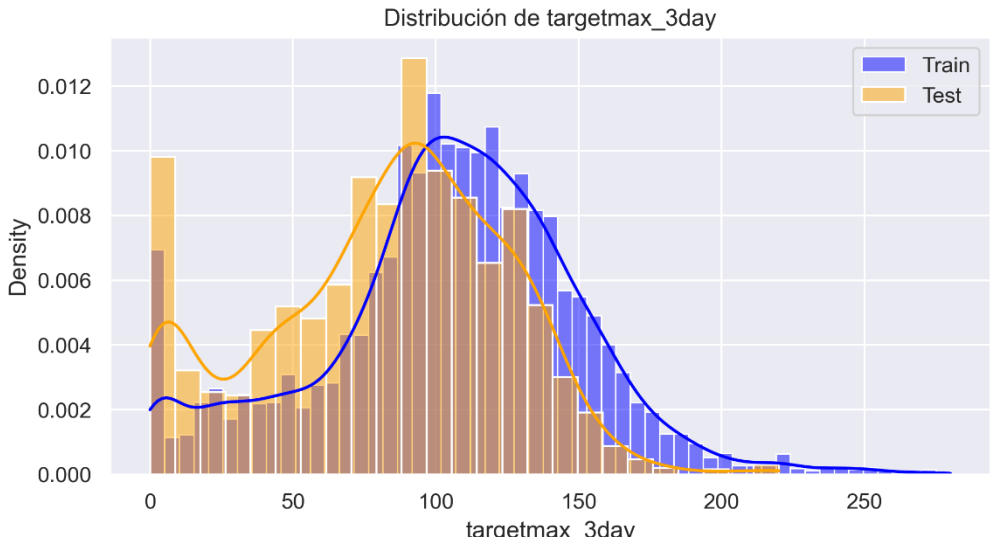

Figure 11. Targetmax distribution

While there is a minor mismatch between the test and train data, either along the target axis or the density axis, the overall shape of the distributions is remarkably similar. This suggests that the underlying patterns in the data are consistent across the two subsets, which is critical for ensuring the model's ability to generalize.

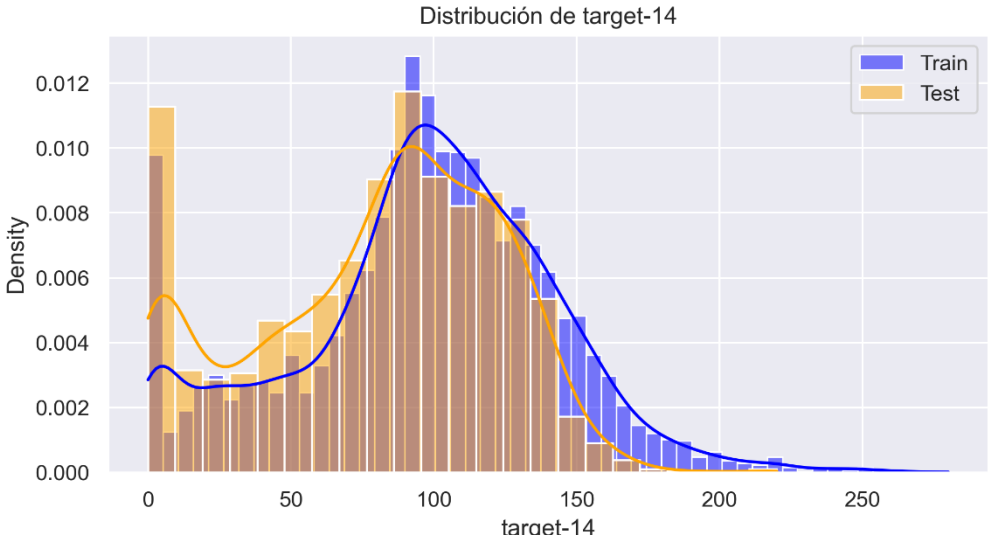

Figure 12. Target-14 distribution

The observed mismatch can be attributed to the temporal nature of the data. Since each group corresponds to different months, it is reasonable to expect slight variations in the environment or conditions under which the data were collected. These temporal factors could influence the exact values and densities of the feature but do not appear to drastically alter its overall distribution.

In the next figures, the values shown seem to be very close to 1, with narrow whiskers in the box plots, indicating minimal variance across the cross-validation folds.

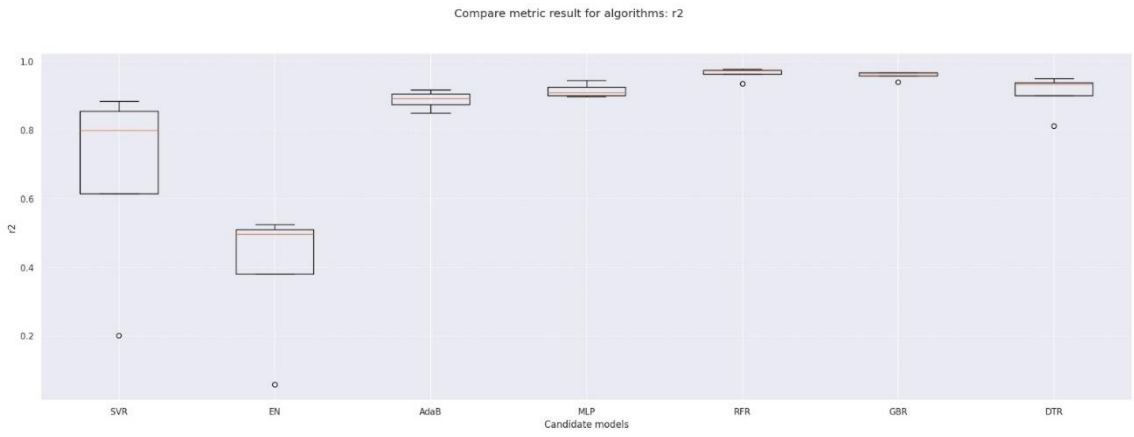

Figure 13. Final model R^2 boxplots

These results highlight the consistency and reliability of the models tested. Among the evaluated models, the Random Forest Regressor (RFR), Gradient Boosting Regressor (GBR), and Decision Tree Regressor (DTR) seam to be the top performers, being excellent in both the R^2 metric and the RMSE.

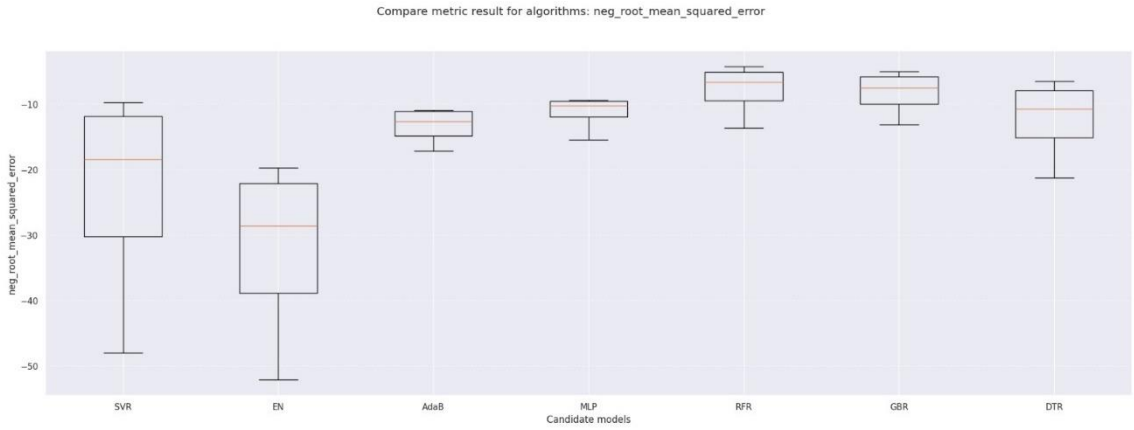

Figure 14. Final model RMSE boxplots

Finaly, the Random Forest Regressor (RFR) was selected as the final model. This decision was based on its beater performance metrics:

- It exhibited the highest mean R^2, indicating the best accuracy and a small positive bias, with a clear tendency toward the ideal value of 1.

- The RMSE for the RFR model showed the smallest mean value and variance, meaning its lees likely to have different results with minimal prediction error.

These results confirm that the RFR model not only achieved the most accurate predictions but also demonstrated high stability and reliability, making it the optimal choice for the task at hand. These results emphasize the critical importance of having meaningful data with strong correlations between features. The absence of such relationships severely limits the model's capacity to make accurate and reliable predictions.

In the final step of evaluation, the actual test values (ytest) were plotted against the predicted values (ypredicted). The results were visually very satisfactory, showing a strong alignment between the two. Even in less common situations or edge cases, the predicted values closely matched the test values, demonstrating the model's ability to generalize well within the dataset.

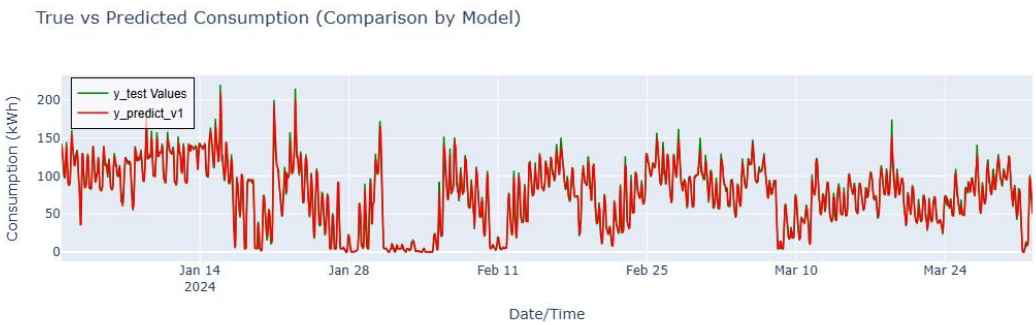

Figure 15. Target test and prediction values

However, this strong performance raises an important question regarding the model's adaptability to environments beyond the scope of the current dataset. While the results indicate excellent accuracy and reliability within the given data, it is critical to assess whether the model can maintain this level of performance when exposed to new, unseen scenarios or datasets with different distributions.

Future steps may include testing the model on out-of-sample data or datasets from different temporal or environmental conditions to evaluate its robustness and capacity to generalize effectively.

4. UNSUPERVISED LEARNING

Unsupervised learning is mainly used to find patterns in an unlabelled dataset. Despite supervised learning, where the goal is to predict a specific output variable (such as determining whether a person has diabetes or not), unsupervised learning focuses on analyzing and understanding input data without predefined labels. The objective is to uncover hidden patterns or clusters within the data, which can provide valuable insights and potentially serve as a foundation for future supervised learning models.

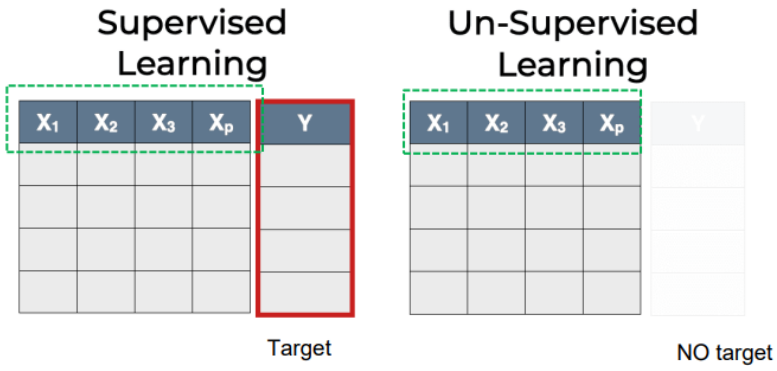

Figure 16. Main differentiation between Supervised and Unsupervised Learning

Since the objective is to find patterns in the input data, the algorithm used will process all this raw data and label it into clusters. In order to achieve those patterns or clusters, a clustering technique from different algorithms which are available for data scientists is needed.

4.1. Clustering

The objective of clustering is to group elements based on their characteristics, more to divide the input data with no labels into groups that the algorithm detects that are similar to each other. It is very useful, as mentioned before, to understand the data and organize it to finally find labels in it.

The K-means and Hierarchical Clustering algorithms are used to implement the clustering. These algorithms, in the first place, need to have information about the optimal number of clusters, thus it is clear that each row of the input data could be a group itself, but from a statistical point of view would make no sense.

In order to assign a number of clusters, it is used the Within-Cluster Sum of Squares, which defines the compactness of clusters. As each cluster has a centroid, it calculates the total squared distances between each data point and the centroid, so for low values of WCSS, the clusters will tend to be more compact.

The elbow method takes advantage of this calculation, and through an easy *for loop*, the value of each WCSS with the *k* centroid is calculated, so it is measured by the compactness of each cluster for *k* number of centroids (clusters). The elbow method gets its name from its characteristic shape on a 2D plot, where the within-cluster sum of squares (WCSS) is plotted against the number of clusters. This method allows data scientists to visually identify the point where the variance begins to decrease at a slower rate as the number of clusters increases, indicating the optimal number of clusters.

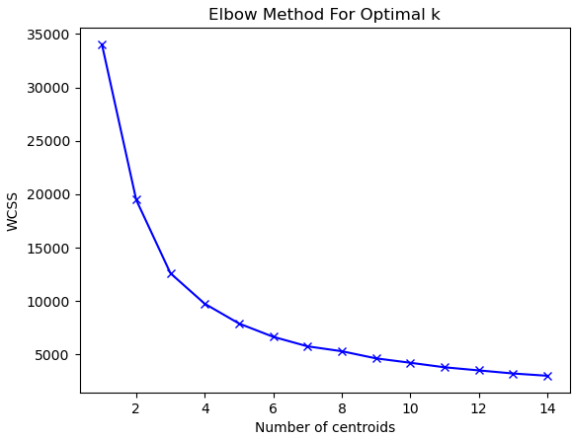

Figure 17. Elbow Method for the optimal k

In this example, with the number of clusters set to 3, more or less the optimal point is found; with a large number of clusters, the sum of squared distance does not decrease as it does before, creating more or less an elbow shape. The number of clusters is essential to perform a good clustering model.

4.2. Methodology

Unsupervised learning techniques such as Hierarchical Clustering and K-Means were employed to identify electricity price patterns. Hierarchical clustering creates a tree-like structure that helps visualize clusters, whereas K-Means assigns data points to predefined k clusters based on their distance to the centroids. The optimal number of clusters was determined using the Elbow Method, which evaluates within-cluster variance, and the Silhouette Coefficient, which measures the cohesion and separation of clusters. These methods were used to explore the segmentation of electricity price data based on influencing variables.

The first analysis aims to identify the most significant variables influencing electricity prices, including forecasted electricity demand, renewable generation such as solar, wind, and solar thermal, natural gas prices, and time factors like the month and whether it was a weekend or weekday. The objective was to find groups with clear trends in electricity pricing and analyse how variables like renewable generation and natural gas prices influence these

groups. For instance, renewable energy generation often drives prices lower, while fluctuating natural gas prices tend to create volatility.

Furthermore, a second analysis was performed including a study of daily electricity price curves, where each curve represents the hourly electricity price for a specific day. These curves provided insights into temporal patterns in price variability. Subsequently, K-Means clustering was applied to identify recurring patterns and dependencies between hourly prices.

The specific objectives of this phase were to process and analyse historical electricity market data, implement unsupervised learning methods to identify meaningful patterns, evaluate the segmentation of temporal series and influencing variables, and create a foundation for future supervised models to predict electricity prices.

To complement this analysis, Principal Component Analysis (PCA) was subsequently used to reduce data dimensionality. PCA transformed the dataset into two principal variables that retained the majority of the information, allowing for easier visualization and interpretation of the clusters. This dimensionality reduction facilitated the identification of key patterns in the data and enhanced the interpretability of results.

4.3. Results

4.3.1. Bivariate clustering of electricity prices and influencing variables

In this section, the results and analysis of the target variable, electricity price, are examined in relation to other significant variables from the dataset, including forecasted demand, solar generation, wind generation, solar thermal generation, natural gas prices, the month, and whether it was a weekend or weekday. The analysis employed unsupervised learning techniques to explore the relationship between electricity prices and each influencing variable in a two-dimensional space. This approach enabled the visualization of clusters, making it possible to observe distinct groups formed by the interaction of electricity prices with these factors.

4.3.1.1. Hierarchical clustering

In this analysis, hierarchical clustering is applied, specifically focusing on the agglomerative clustering method. Agglomerative clustering is a bottom-up approach to hierarchical clustering where each data point initially starts as its own individual cluster. These clusters are then iteratively merged based on their similarity until a single cluster encompassing all data points is formed or a predefined number of clusters is achieved.

The data used in this analysis were scaled using the Standard Scaler method. Standard Scaler transforms the data by removing the mean and scaling it to unit variance, effectively

centering the data around zero with a standard deviation of one. This ensures that all variables are on the same scale, which is particularly important for clustering algorithms like Agglomerative Clustering, as they rely on distance metrics (Euclidean distance) that can be distorted by differences in variable scales.

One of the key benefits of using Standard Scaler is the reduction in computational time, as standardized data allow algorithms to converge faster and more efficiently. Additionally, it improves the accuracy of clustering results by ensuring that no single variable dominates the clustering process due to larger scales.

Next, the results of the clustering method are visualized for configurations with 2, 3, and 4 clusters. The variables represented in the plots include electricity price on the x-axis and various influencing factors on the y-axis, such as demand, wind generation, solar generation, solar thermal generation, and natural gas prices.

These visualizations allow for the observation of how the data points are grouped under different clustering configurations. Each chart highlights the relationship between electricity prices and one of the influencing variables, making it possible to discern how different factors contribute to the formation of distinct groups.

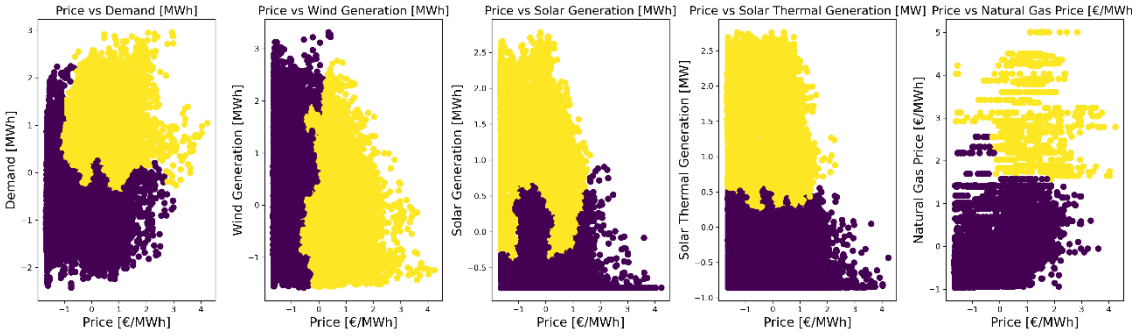

Figure 18. Agglomerative analysis results for electricity price and influencing variables for 2 clusters

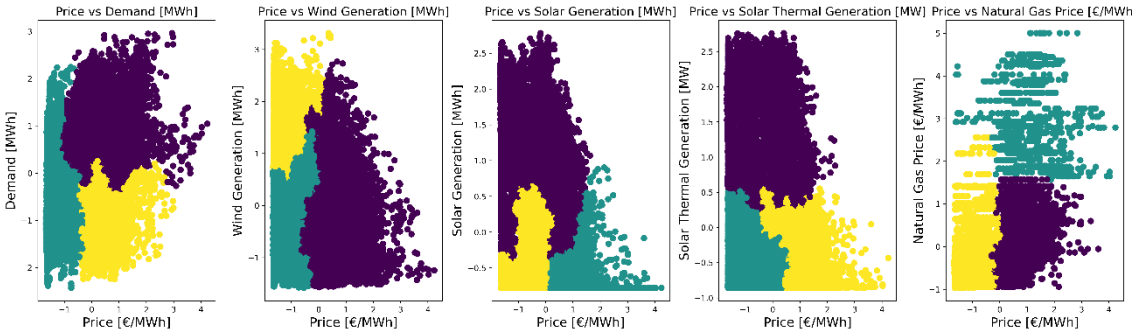

Figure 19. Agglomerative analysis results for electricity price and influencing variables for 3 clusters

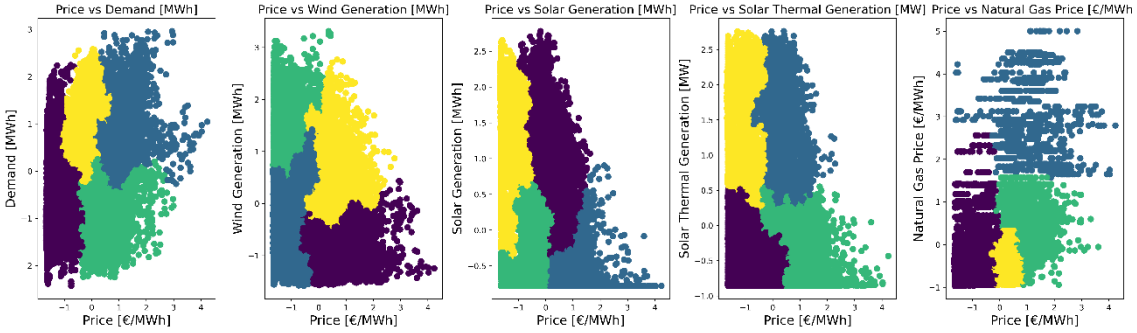

Figure 20. Agglomerative analysis results for electricity price and influencing variables for 4 clusters

The visualizations clearly show that when electricity prices are significantly elevated, all renewable generation sources, including wind, solar, and solar thermal, are at their lowest levels of generation. This suggests that reduced renewable energy availability is associated with higher electricity prices. Similarly, a notable pattern emerges with natural gas prices and electricity demand. When either natural gas prices or demand is high, electricity prices also tend to rise. This indicates a strong positive correlation between these variables and electricity prices, where an increase in either natural gas prices or demand appears to drive up electricity costs. These relationships highlight the critical influence of both renewable generation and fossil fuel-based energy costs on electricity pricing.

Subsequently, the results of the clustering method are also visualized for configurations with 2, 3, and 4 clusters, considering additional categorical variables such as the month of the year and whether it is a weekend or not. By incorporating these time-related variables, the analysis aims to uncover potential seasonal patterns or differences in electricity prices based on the day type, providing further insights into how temporal factors influence pricing dynamics.

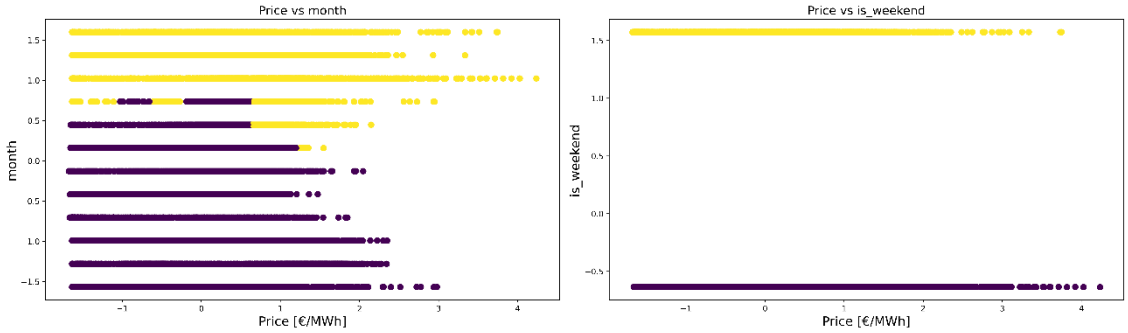

Figure 21. Agglomerative analysis results for electricity price and time-related variables (month and weekend) for 2 clusters

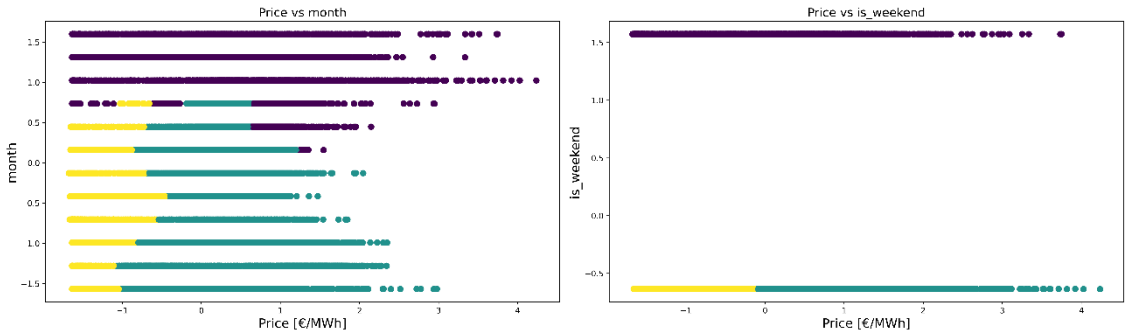

Figure 22. Agglomerative analysis results for electricity price and time-related variables (month and weekend) for 3 clusters

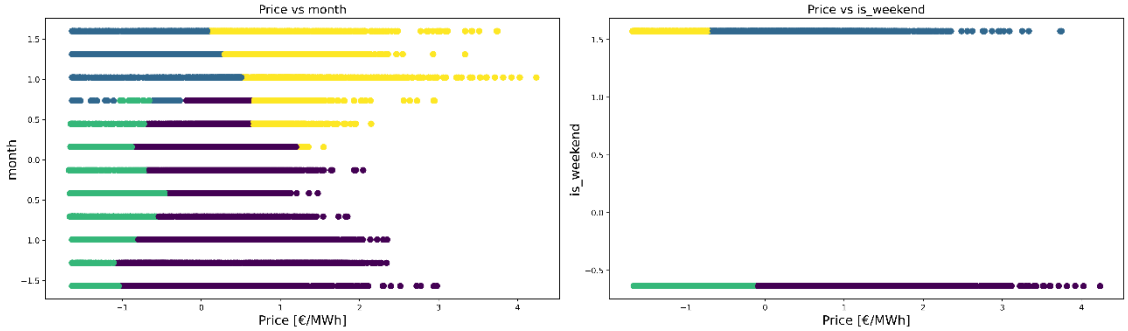

Figure 23. Agglomerative analysis results for electricity price and time-related variables (month and weekend) for 4 clusters

The analysis reveals that electricity prices are noticeably lower during the summer months compared to the latter part of the year. This trend could be explained by increased solar energy generation during summer and the reduced heating demand. On the other hand, electricity prices peak in October, with the highest values. This increase might be driven by higher energy demand for heating and reduced renewable energy availability, particularly solar generation. Additionally, weekends display slightly less price variation, which could reflect a decrease in industrial energy consumption compared to weekdays.

4.3.1.2. K-Means

In order to perform a more extensive algorithm, the K-mean is also implemented in the analysis. This algorithm divides the input data into *k* clusters, or groups, which each one has its own centroid or k-center.

The algorithm, once it is given the information about the number of clusters, selects randomly where the centroids are placed. In python, it is required to select a random state since it cannot alone decide randomly an iteration. Later it starts to assign each data point to their closest centroid, which will form the predefined K clusters.

The centroids now need to be in a different position, in order to reduce as maximum as

possible the variance of the cluster. That is why in every iteration the centroids change their position and reduce the variance until they do not change anymore while recalculating the variance itself.

To analyze the relationship and find patterns between two different variables of the dataframe, it could be interesting to make a scatterplot of these two different variables in a 2D frame where every dot represents one hour over the two years dataframe.

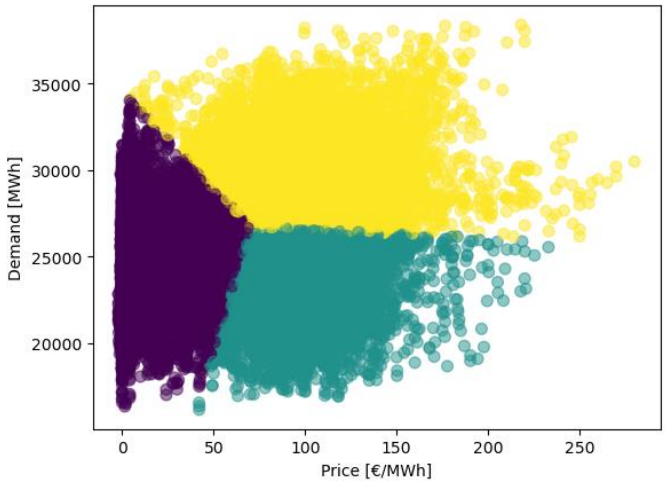

Figure 24. Patterns found on Demand vs Price of electricity

The K-means algorithm got three different groups on the demand over the price of electricity. The purple group represents hours where the prices are cheaper and the demand is also kind of low; it could represent instants where the renewable energy generation is higher and the demand is low so the prices tend to be cheaper. The green group could represent a case where even though there is low demand, the generation could proceed from gas natural or combined cycle due to lack of solar and wind generation, leading to high prices. The yellow group agglomerates more or less all the high energy demand where most of the prices are between 60 - 150 €/kWh, including scenarios where the generation source could vary.

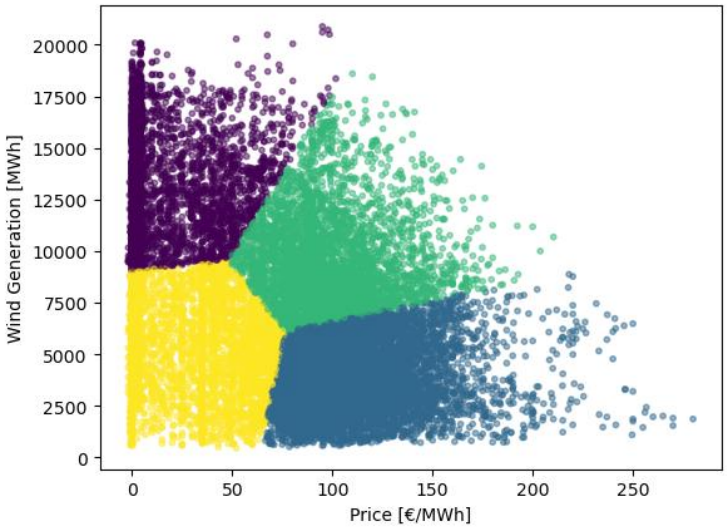

Figure 25. Patterns found on Wind generation vs Price of electricity

The wind generation is divided into 4 different blocks; the purple one represents cheaper prices and high wind generation, the yellow one represents also cheaper prices but low wind generation, the green one represents kind of medium and high wind generation but the prices are quite high than the purple one, and the blue one represents low wind generation and high prices.

The purple and blue cases are completely opposite and are quite understandable, when there is wind generation the prices tend to be lower and where there is a lack of it, the prices tend to increase. The yellow scenario could indicate, for instance, high solar generation (sunlight hours). The green cluster could be representing hours of high demand and lack of solar generation, so only generating from wind sources is not enough to reduce the electricity price.

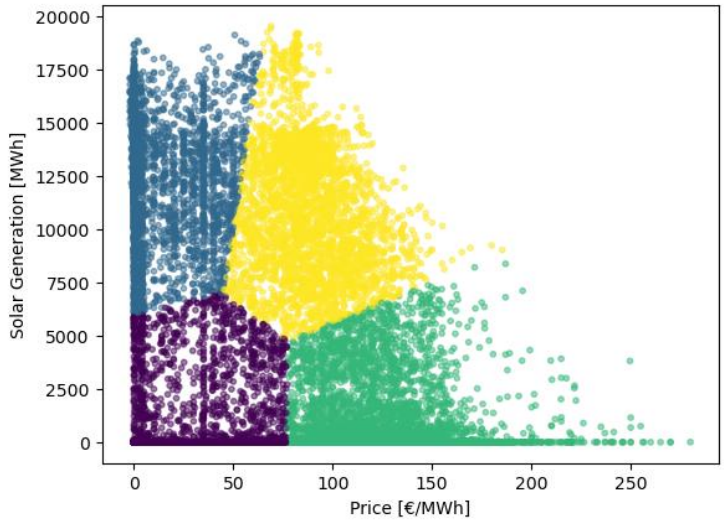

Figure 26. Patterns found on Solar generation vs Price of electricity

The solar photovoltaic generation is also divided into four clusters and also represents the same scenarios described in wind generation; these two sources are complementary. Some differences appear in the density of these clusters, where the cheap prices over the lack of generation tend to be less common than in wind sources.

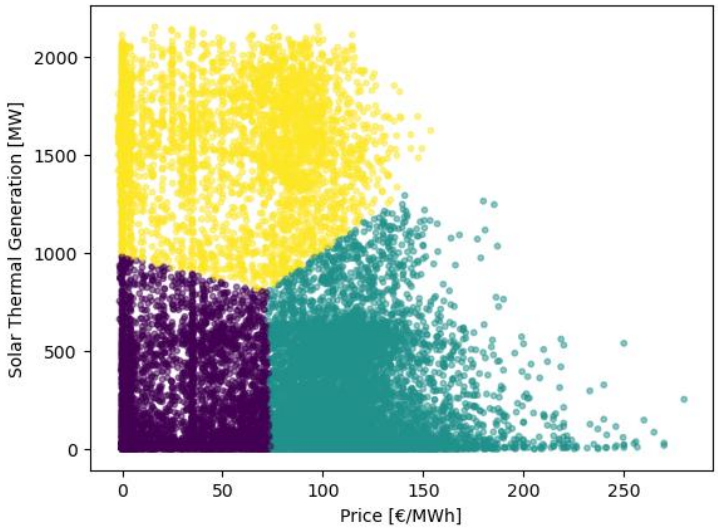

Figure 27. Patterns found on Solar thermal generation vs Price of electricity

Solar thermal energy is a technology that also requires a solar source to generate energy, but it requires more OPEX costs than photovoltaic generation. In the electricity market, as solar and wind enter with practically 0 €/kWh in the daily electricity market, as they always want to enter, solar thermal needs to join at a higher price.

The evolution is quite similar to photovoltaic generation, when there is a high source, the price does not surpass 150 €/kWh, considering different scenarios where the wind resource is low so the price increases or the demand is high as well. The green and purple clusters reflect all the prices where the resource is low.

Other things done in the Hierarchical Clustering part where there have been defined patterns over the price on the weekend and on different seasons affect also the price in this clustering scenario but it has already been defined in the previous chapter.

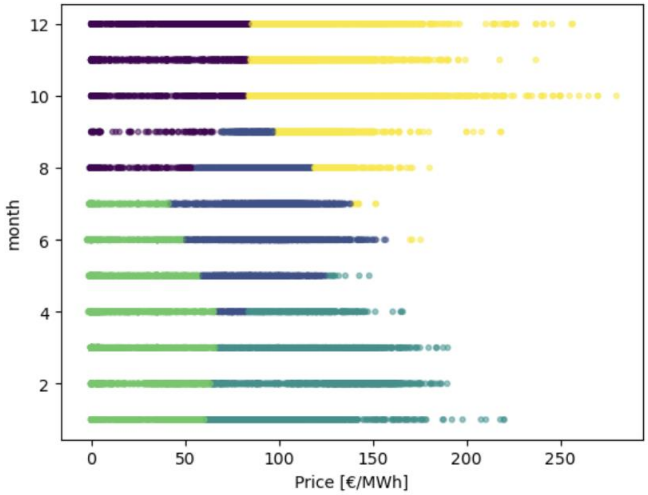

Figure 28. Scatter plot of the electricity price per each month

4.3.1.3. Principal Component Analysis (PCA)

A Principal Component Analysis (PCA) was conducted to reduce the dataset to two principal component variables. The scatter plot visualizes the results, with each point representing the data projected onto the two principal components. These components encapsulate the most critical information from the original variables, facilitating further analysis and enabling easier interpretation of the underlying patterns.

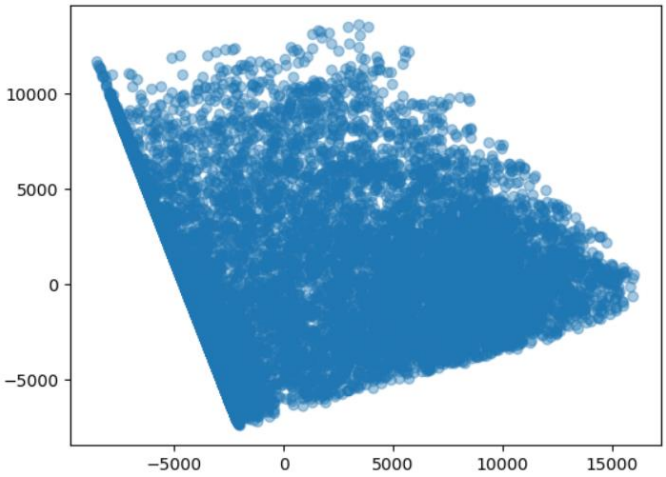

Figure 29. Scatter Plot of Data Projected onto Two Principal Components after PCA

Using the Elbow Method, it was determined that the optimal number of clusters is 3. The resulting graph illustrates the clustering of the data projected onto the two principal components (PC1 and PC2). Each cluster is represented by a distinct color, providing a clear visualization of the groupings identified through the clustering algorithm.

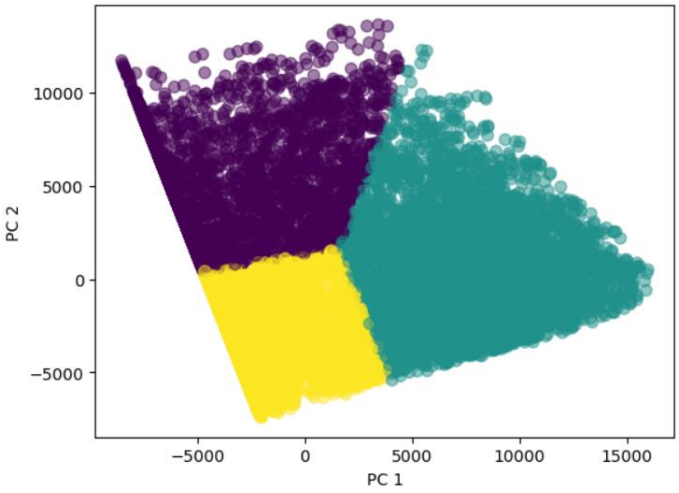

Figure 30. Scatter Plot of Data Projected onto Two Principal Components after PCA with 3 Clusters

The clusters identified during the PCA analysis are now integrated into the original dataset, assigning a cluster label to each row of data. This allows each data point in the dataset to be grouped according to the patterns identified in the PCA clustering. Subsequently, various plots are generated, showing electricity price versus all other variables in the dataset, with the assigned clusters visually represented.

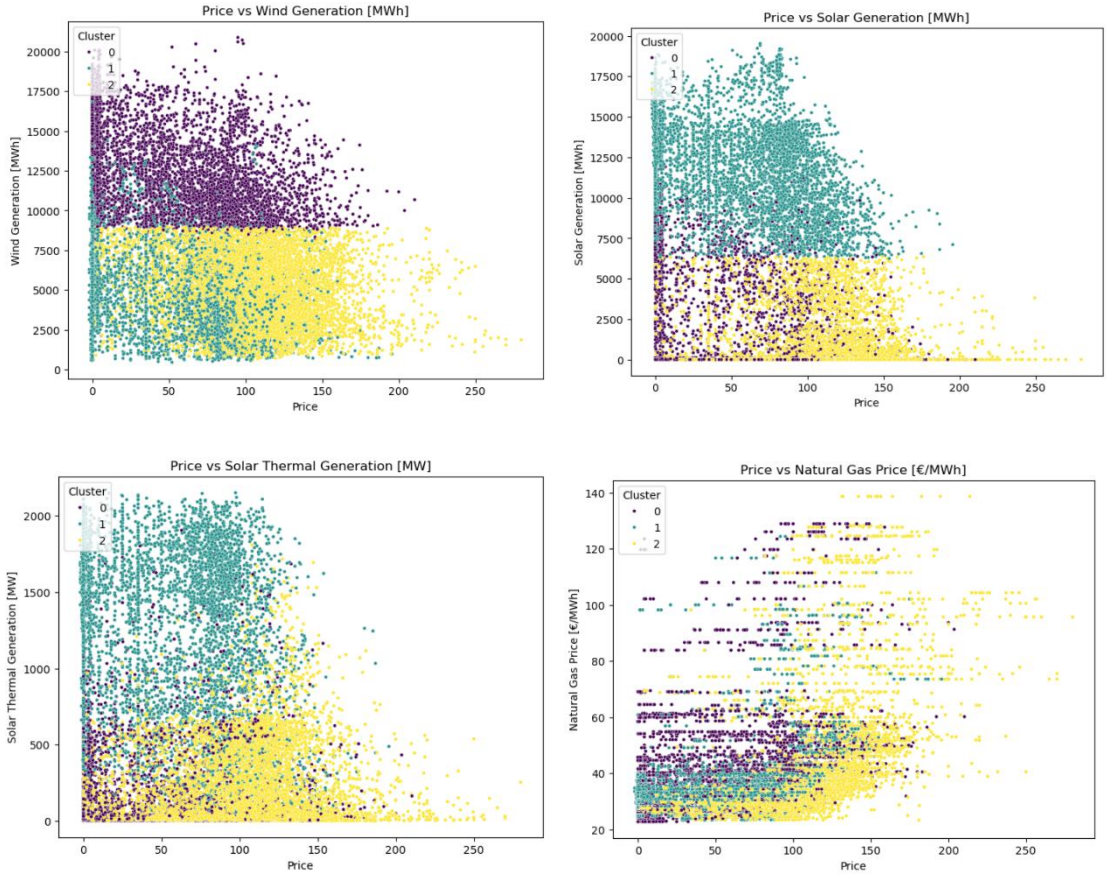

Figure 31. Cluster analysis of electricity price vs. renewable generation and natural gas price using PCA-defined groups

ETSEIB

The analysis of the clusters reveals distinct patterns when comparing wind generation with solar and solar thermal generation, highlighting the differing impacts of these renewable sources on electricity prices across the identified groups.

For wind generation, the purple cluster corresponds to scenarios where wind production is high, and electricity prices are low or moderate. This indicates that an abundant wind energy supply significantly contributes to reducing electricity prices. The blue cluster in wind generation represents moderate wind production and slightly higher electricity prices, reflecting conditions where wind energy alone may not be sufficient to maintain low prices but still plays a stabilizing role. Finally, the yellow cluster represents high electricity prices with low wind generation, demonstrating that insufficient wind energy leads to greater reliance on more expensive energy sources, thereby driving up electricity costs.

In contrast, for solar and solar thermal generation, the purple cluster corresponds to low solar production with low to moderate electricity prices. This pattern emerges because solar generation is often complemented by other renewable sources, like wind, which keeps prices low even when solar output is minimal. The blue cluster, on the other hand, represents high solar generation associated with low to moderate electricity prices, underscoring the significant role of solar energy in reducing electricity costs. Similar to wind generation, the yellow cluster corresponds to high electricity prices when solar and solar thermal generation are both low, reinforcing the reliance on non-renewable sources in such scenarios.

These patterns highlight an important insight: if one renewable energy source, either wind or solar, is producing at high levels, electricity prices tend to decrease or stabilize. Conversely, when both renewable sources are generating at low levels (as represented by the yellow cluster), electricity prices rise significantly due to increased reliance on costlier energy sources. This underscores the complementary role of wind and solar energy in moderating electricity prices and the critical need for diversified renewable energy generation.

Lastly, the clustering in the Electricity Price vs Natural Gas Price plot does not reveal particularly distinct patterns or meaningful separations between the groups. The purple cluster appears to dominate at lower electricity prices and spans a wide range of natural gas prices, suggesting that low electricity prices can occur across varying natural gas price levels. The blue cluster occupies a mid-range region for electricity prices and natural gas prices, showing no strong or consistent relationship between the two variables. The yellow cluster, associated with high electricity prices, is spread across higher natural gas price ranges but overlaps significantly with the other clusters.

4.3.2. Analysis of daily electricity price curves and temporal patterns using K-Means clustering

The graph illustrates the daily electricity price curves, where each curve represents the hourly electricity price for a specific day. The figure displays a total of 730 lines, representing the hourly electricity price curves for two years of data. The x-axis corresponds to the 24 hours of the day, while the y-axis indicates the electricity price in €/MWh. Each individual curve reflects the temporal price variation across different hours for a particular day.

The visualization provides insights into daily patterns, highlighting how electricity prices fluctuate within a day and vary across multiple days. Peaks typically represent hours of high demand, such as morning or evening, while troughs indicate hours of lower demand, such as late at night. This temporal analysis is useful for identifying recurring trends, seasonal variability, or anomalies in the electricity market.

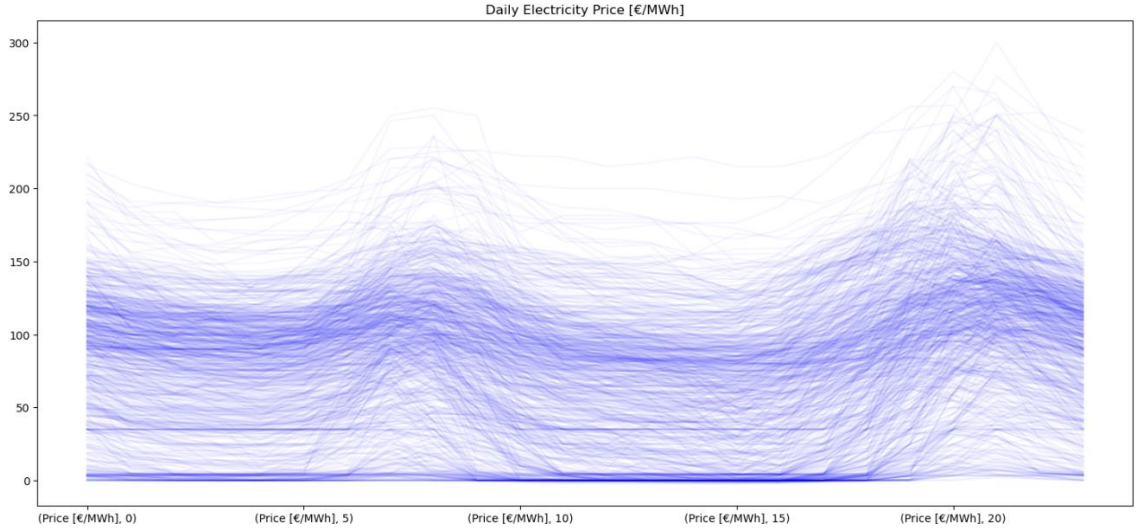

Figure 32. Hourly electricity price curves over two years (730 daily curves)

The Silhouette Coefficient Method was used to determine the optimal number of clusters. This method evaluates the quality of clustering by measuring how well-separated and cohesive the clusters are. In this case, the analysis revealed that the optimal number of clusters is two, as the Silhouette Coefficient reaches its highest value at this point. This indicates that dividing the data into two clusters provides the most distinct and well-defined grouping based on the data structure.

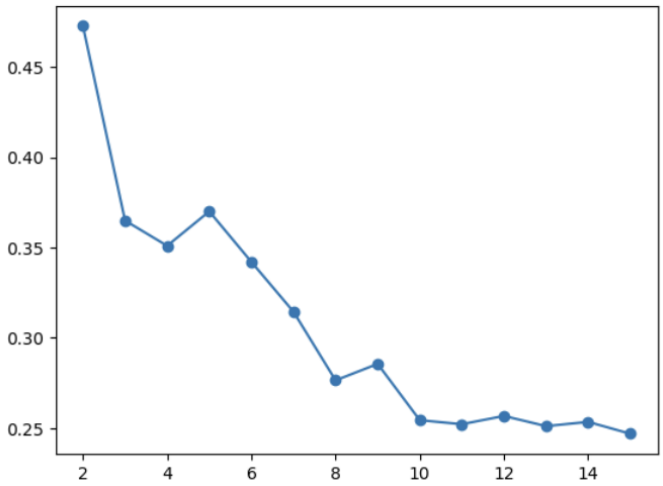

Figure 33. Optimal number of clusters determined by the silhouette coefficient

Using the K-Means clustering method with two groups, the daily electricity price curves were classified into two distinct clusters, as shown in the graph. The red cluster corresponds to days with consistently higher electricity prices throughout the day, with a noticeable peak during the late morning and evening hours. This group likely reflects periods of higher demand or reduced availability of renewable energy, which drives prices up.

In contrast, the blue cluster represents days with lower electricity prices, showing much smaller peaks and generally more stable price patterns throughout the day. These days are likely characterized by sufficient renewable energy generation or lower overall electricity demand, which keeps prices moderate.

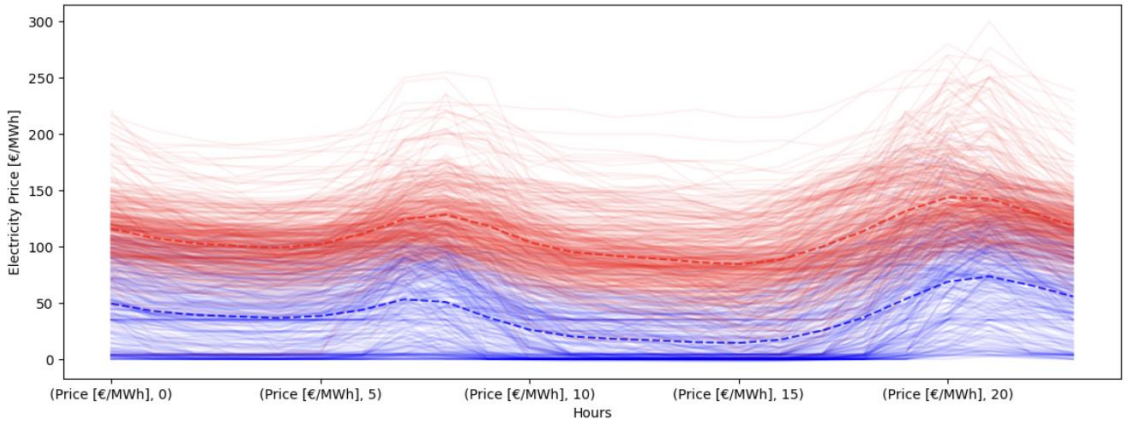

Figure 34. Daily electricity price curves clustered using K-Means (2 groups)

The process involves first applying Principal Component Analysis (PCA) to reduce the dimensionality of the dataset, projecting it onto two principal components that capture the maximum variance. After this dimensionality reduction, the groups previously assigned by the K-Means clustering algorithm are visualized on the PCA components.

The resulting scatter plot highlights the distinctiveness of the clusters, with each point colored according to its K-Means group. The clear separation between the clusters in the PCA space demonstrates that the K-Means clustering successfully identified meaningful groupings. This approach validates the robustness of the clustering results, as it shows that the groups assigned by K-Means align well with the main patterns captured in the reduced dimensions by PCA.

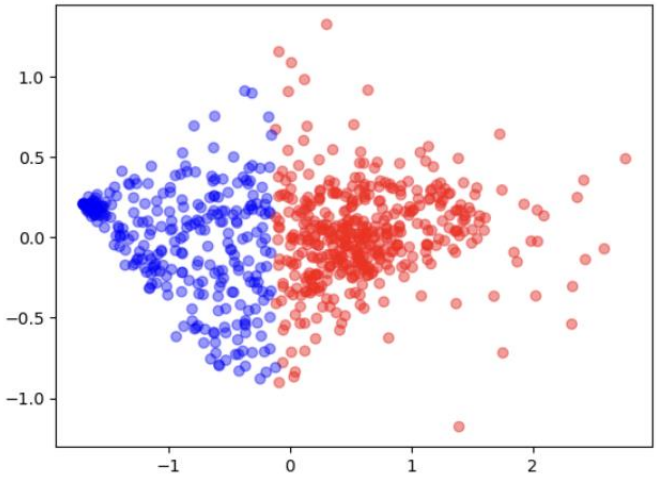

Figure 35. Validated results with dimensionality reduction (PCA) with two K-Means clusters

Additionally, an analysis was performed to determine the optimal number of clusters using the Elbow Method. In this case, the analysis indicates that the optimal number of clusters is three, as the elbow of the curve is visible at this point. This result provides further validation for the grouping structure in the dataset.

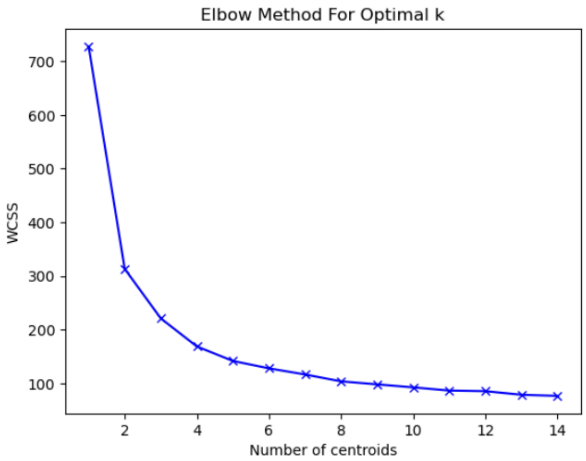

Figure 36. Elbow Method for optimal k

The K-Means clustering method was applied with three groups, resulting in the red, green, and blue clusters shown in the graph. The red cluster represents days with significantly higher electricity prices, with noticeable peaks at specific hours. These elevated prices

could be due to increased demand during colder months or limited renewable energy generation. The presence of outliers in this cluster suggests occasional extreme events, such as unexpected demand surges or supply disruptions.

The green cluster corresponds to days with moderate electricity prices, reflecting stable conditions with balanced demand and supply. Finally, the blue cluster represents days with consistently lower electricity prices, likely associated with high renewable energy availability or reduced demand. The clear separation between these groups highlights the variability in electricity price behaviour across different scenarios.

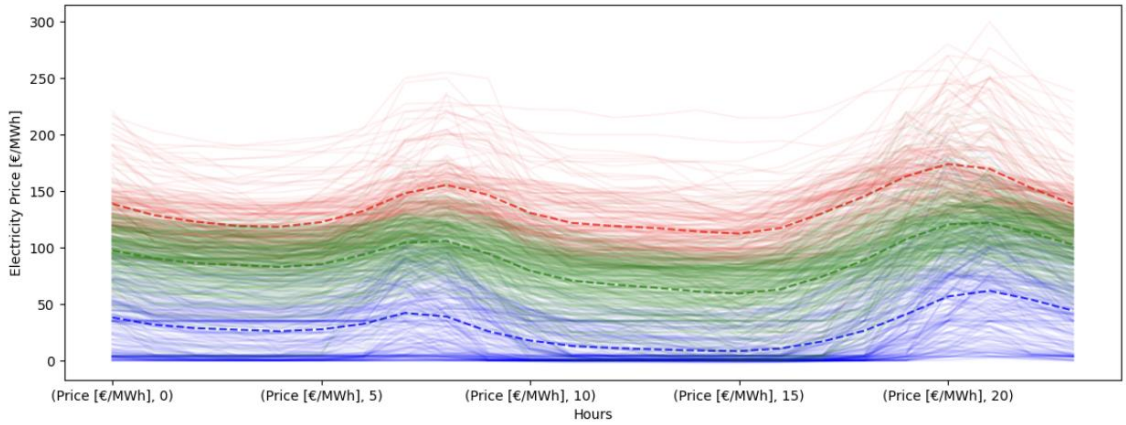

Figure 37. Daily electricity price curves clustered using K-Means (3 groups)

The final validation using PCA clearly distinguishes the three clusters in the reduced two-dimensional space. This separation confirms the robustness of the K-Means clustering, highlighting well-defined groupings in the dataset.

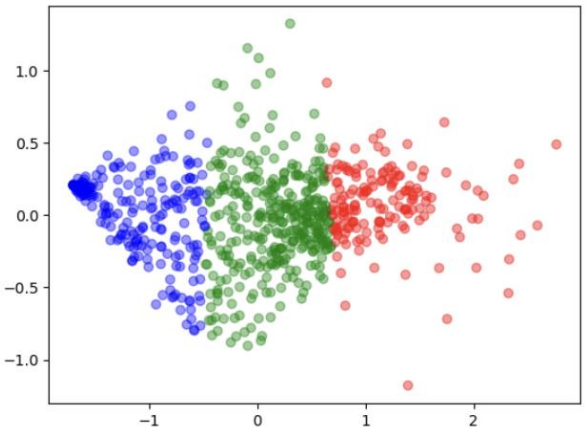

Figure 38. Validated results with dimensionality reduction (PCA) with three K-Means clusters

5. CONCLUSIONS

The supervised learning approach in this project demonstrated the critical importance of selecting relevant features and understanding their relationships with the target variable. Through feature engineering and refinement, the model successfully identified the most influential factors, enhancing predictive performance. Among the models tested, the Random Forest Regressor emerged as the most accurate and reliable, effectively leveraging meaningful patterns in the data. Evaluation metrics, including R^2 and RMSE, confirmed its robustness, while temporal and external variables, such as renewable generation and natural gas prices, proved crucial in capturing market dynamics. This highlights the importance of robust data preprocessing and feature selection in optimizing machine learning models for complex systems like electricity markets. Future work could integrate additional variables or test the model on diverse datasets to assess its adaptability.

The unsupervised learning analysis revealed distinct electricity price patterns influenced by renewable generation and natural gas prices. K-Means and hierarchical clustering successfully grouped data into meaningful clusters, with PCA enhancing interpretability by reducing dimensionality. Clustering identified three main groups: one with high renewable generation and low prices, another with low generation and high prices, and a third representing low prices and generation during specific scenarios, such as weekends, holidays, or summer months. These clusters provide a clearer understanding of market dynamics and could support supervised models by creating new labels to improve electricity price predictions.

Natural gas prices remain a key driver of electricity pricing, particularly when renewable output is low and combined cycle plants are required to meet demand. Spain’s dependence on natural gas imports highlights the importance of maintaining strong international relationships with suppliers like Algeria and Russia. However, geopolitical tensions, such as the EU's support for Ukraine, complicate these relationships and pose challenges to energy security. Balancing these dynamics is essential for ensuring a stable and affordable energy supply.

6. BIBLIOGRAPHY

[1] S. Raschka, Y. Liu, and V. Mirjalili, Machine Learning with PyTorch and Scikit-Learn. Packt, 2022.

[2] L. Igual and S. Seguí, Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications. Springer, 2017.

[3] M. Nazari-Heris, S. Padmanaban, J. P. S. Catalão, M. Hosseini, M. Shakeri, and A. M. Abdel Aleem, Application of Machine Learning and Deep Learning Methods to Power System Problems. Springer, 2021.

[4] J. Brownlee, Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-to-End. Machine Learning Mastery, 2016.