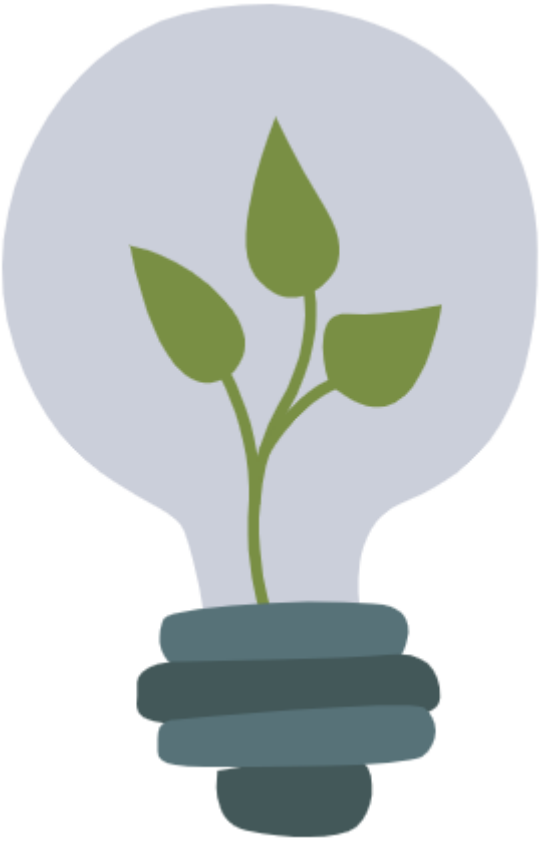
Synthesis Project - Report

Team 3 - WATTWise



Joaquín Arenas

Martina Carretta

Mariona Pla

Ares Sellart

Index of contents:

Introduction:

In the context of increasing energy demands and the imperative for sustainability, the WATTWise project emerged as a crucial initiative aimed at fostering wise management of energy consumption. This project was specifically designed to assist the energy manager of the UAB in gaining a clearer understanding of energy usage patterns and identifying strategies to reduce energy consumption.

The University Autònoma de Barcelona (UAB) has a big main campus in Bellaterra. This project focused on optimizing the energy consumption of building Q. Building Q is the Escola d'Enginyeria faculty placed on the far East side close to the science building. Given that building Q is part of a university and has a lot of technological equipment, the energy consumption is very elevated. The building features computer rooms, material rooms with machines available for 3D printing and various types of crafts, air conditioning for the entire building, projectors, and screens, among other things.

To achieve wise management of energy consumption, the project employed sophisticated data handling processes to manage sensor data, occupation data, energy consumption data, and PV production data. Machine learning models were trained to predict future energy consumption and occupancy levels, providing actionable insights that can inform energy-saving measures.

By understanding how energy is consumed and how buildings are occupied at different times, the energy manager can make informed decisions to reduce waste and improve operational efficiency. This comprehensive approach not only contributes to significant energy savings but also supports the university's commitment to sustainability.

Furthermore, the project included the development of a user-friendly website that displays real-time data and predictions, making it easier for the energy manager and other stakeholders to access and interpret the information. This website also features a recommendation system designed to optimize classroom usage, further aiding in the reduction of energy consumption.

In summary, the WATTWise project represented a significant step forward in energy management for the UAB, leveraging cutting-edge technology to support sustainability and operational efficiency.

Proposal of the solution:

The WATTWise project was guided by several key objectives. First and foremost understanding data patterns. One of the fundamental goals was to delve into the sensor data, occupation data, energy consumption data, and PV production data to uncover underlying patterns and correlations. This understanding was essential for making informed decisions about energy management.

Working with real-time data was another of the main objectives alongside predicting energy consumption. By leveraging historical data and machine learning models, the project aimed to

accurately predict future energy consumption. These predictions could help in planning and optimizing energy usage, thereby reducing waste and improving efficiency. Another critical aspect of the project was to forecast faculty occupation levels. Understanding how different areas are utilized at various times can assist in aligning energy usage with actual occupancy, ensuring that energy is not wasted on unoccupied spaces.

Another key objective of the project was to predict the production of the solar panels installed at the faculty, taking into account meteorological data. Accurate predictions of solar panel output are essential for integrating renewable energy sources effectively. By analyzing weather patterns and their impact on solar energy production, the project aimed to optimize the use of solar power, reduce reliance on non-renewable energy sources, and enhance overall energy efficiency.

As part of the final product, an energy-saving recommender system had a big focus. An innovative feature of the project was the creation of a recommendation system aimed at optimizing classroom usage to save energy. This system suggests actions based on the real-time sensor data such as opening the window, closing the lights…

The environment hosting the collection of previously mentioned objectives is a webpage for information display. The project included the development of a comprehensive website that displays real-time data from sensors, predictions generated by machine learning models, and alerts or warning messages. This website serves as a user-friendly interface for stakeholders to access and understand energy consumption patterns and forecasts.

Regarding the AI methodology, the initial plan was to explore several models, including Linear Regression, Decision Trees, Random Forests, Gradient Boosting Machines (GBM), Convolutional Neural Networks (CNNs), and Multi-Layer Neural Networks. Although the subject of deep learning provided some ideas, the practical knowledge to implement these models was limited. Consequently, only machine learning techniques were utilized.

To achieve the objectives, various approaches will be considered; rule-based methods as a baseline and machine learning models to go a bit further.

Ethical issues:

The ethical aspects of smart energy management revolved around promoting key values such as sustainability, energy efficiency, privacy, transparency, accountability, well-being, educational value, accessibility, and technical robustness. Although explicit discussions on values were absent, a consensus on the project's objectives, particularly its environmental goals, underscores a shared commitment to sustainability and privacy.

The project aimed for significant energy savings, reduced carbon footprint, and improved environmental conditions, thereby enhancing well-being, and educational value; and serving as a scalable model for other institutions. Stakeholders, including students, faculty, administration, and the local community, benefit from improved conditions and reduced costs, with careful measures to avoid discrimination and ensure accessibility.

Implementation of the solution:

The main tasks were divided among the group integrants to ensure efficient progress and coverage of all necessary aspects. Each team member had specific responsibilities corresponding to their expertise or assigned roles within the project.

Additionally, the team engaged in proactive planning, including short-term and long-term scheduling which was materialized on GitHub. Two-week planning ahead would have involved detailed scheduling of activities and deadlines to keep the project on track in the immediate future. This planning allowed the team to anticipate challenges, allocate resources effectively, and adjust strategies as needed.

Furthermore, a general plan for the future was formulated, outlining broader objectives, strategies, and milestones beyond the immediate scope of the project. This long-term plan provided a more open view of the project. Engaging in both short-term and long-term planning ensured the project's success while also laying the groundwork for its ongoing development and expansion.

1. Occupation dataset:

The occupation dataset encompasses data spanning from 2018 to 2024. This dataset includes detailed information on various aspects such as the subjects taught, the number of students in each class, and the duration of each class on a daily basis. To ensure the dataset's usability, several preprocessing steps were implemented:

1. Deduplication: duplicate rows were removed to eliminate redundant information and maintain the dataset's integrity.

2. Hourly splitting and aggregation: classes were divided into hourly segments, and the relevant data was aggregated accordingly to provide a more granular view of class occupancy.

3. Handling missing data: for instances where student information was missing, the dataset was supplemented with an estimated value, set at 60% of the maximum capacity of the respective class.

2. Energy consumption dataset:

The energy consumption dataset provides detailed hourly information regarding the energy usage of the university's faculty buildings. To facilitate a better understanding of overall consumption patterns, the data was initially aggregated into monthly totals. This approach allowed for a comprehensive analysis of energy trends over extended periods. From it, two main topics can be highlighted.

Impact of the COVID-19 pandemic:

A significant reduction in energy consumption was observed starting in March 2020, coinciding with the onset of the COVID-19 pandemic. The closure of the university and the transition to

online classes led to a marked decrease in energy usage. This downward trend persisted even beyond the initial lockdown period, with a modest decrease in energy consumption recorded in the subsequent months. The pandemic-induced changes highlight the substantial impact of operational shifts on energy usage.

Seasonal consumption patterns:

Analyzing the dataset (see *Figure 1*) revealed that August consistently exhibits the lowest energy consumption levels each year. This is primarily due to the reduced need for air conditioning during this month, as most university activities pause for the summer break.

Conversely, an unexpected peak in energy consumption is observed in July, which presents a noteworthy anomaly. Given that the majority of classes typically conclude by this time, this spike in energy usage was intriguing. Several factors may contribute to this phenomenon:

- Summer courses: the operation of summer courses, which require facility use and, consequently, energy.
- Faculty and administrative activities: continued use of buildings by faculty members and administrative staff during the summer.
- Maintenance activities: potential maintenance work conducted during the summer break, necessitating energy use.

Additionally, certain infrastructural constraints contribute to ongoing energy consumption. Air conditioning systems in specific faculty corridors cannot be completely switched off, leading to their continued operation even when classes are not in session. Furthermore, essential machinery, servers, and other equipment remain active throughout July, adding to the overall energy usage.

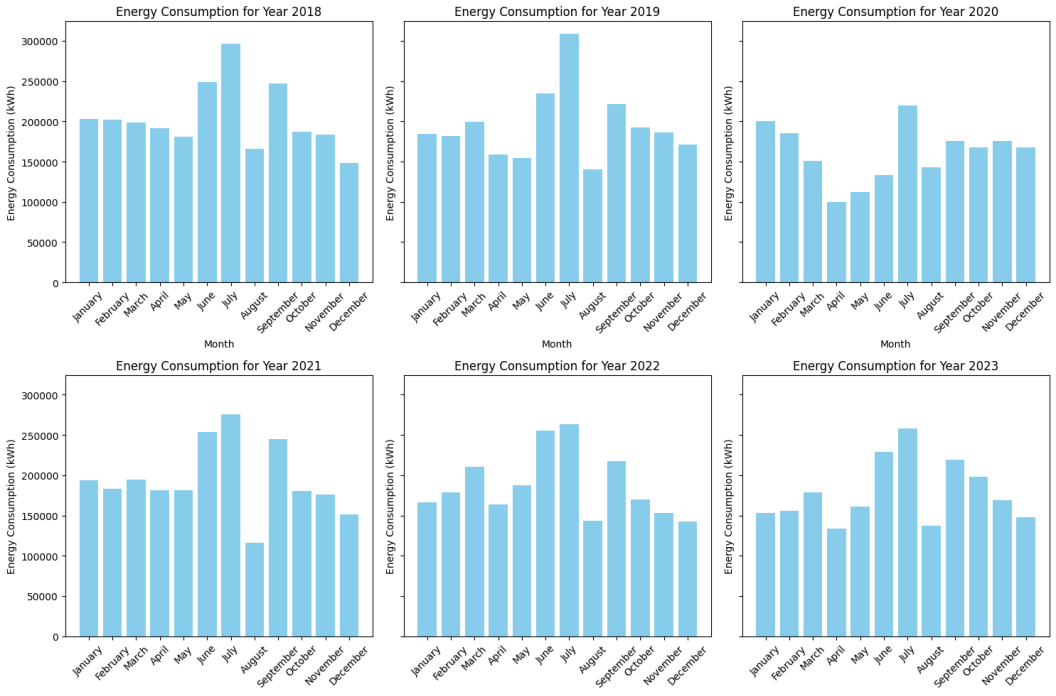


Figure 1: Plot formed by 6 bar histograms each being a year from 2018 to 2023. X axis contains the 12 months labelled and y axis represents the energy consumption in kWh.

3. Sensors dataset:

The sensors dataset was a crucial component of the WATTWise project, providing detailed environmental and operational data from various locations within the university. This dataset encompasses a range of sensor types installed across different rooms and facilities, capturing key metrics that aid in monitoring and managing the university's energy consumption and indoor environment quality.

The following sensors were installed at specific locations within the university:

1. eui-24e124710c408089: OpenLab – Laser room
2. eui-24e124128c147444: Biblioteca de filosofia i lletres
3. eui-24e124128c147500: OpenLab – main room
4. eui-24e124128c147204: DigitalLab
5. eui-24e124128c147499: AudioLab
6. am307-9074: Computer Room
7. q4-1003-7456: Q4-1003
8. eui-24e124128c147446
9. eui-24e124128c147470

The sensors capture a variety of environmental and operational parameters, each providing insights for managing the university's facilities:

- CO_2: Measures the concentration of carbon dioxide in the air, indicating air quality and ventilation efficiency.
- Humidity: Measures the amount of water vapor present in the air, essential for maintaining comfort and preventing mold growth.
- Light Level: Measures the intensity of light in an environment, important for ensuring adequate lighting conditions.
- O3: Measures the concentration of ozone in the air, which can impact respiratory health.
- PM10: Measures particulate matter with a diameter of 10 micrometers or less, indicating air quality and potential health risks.
- PM2.5: Measures particulate matter with a diameter of 2.5 micrometers or less, providing insights into finer air pollutants that can penetrate deep into the lungs.
- Pressure: Measures the atmospheric pressure, which can influence weather conditions and indoor air quality.
- Temperature: Measures the degree of hotness or coldness of an environment, crucial for maintaining comfortable indoor conditions.
- TVOC: Measures the total volatile organic compounds in the air, indicating the presence of chemical pollutants.
- Battery: Measures the remaining energy or charge level of a battery, ensuring the sensors are functioning properly.
- Activity: Measures movement or action within a space, providing data on room occupancy and usage.
- Illumination: Measures the brightness of light in a given space, useful for assessing lighting conditions.

- Infrared: Measures infrared radiation emitted or absorbed by objects, used for detecting heat and movement.
- Infrared and Visible: Measures both infrared and visible light wavelengths, providing comprehensive light data.

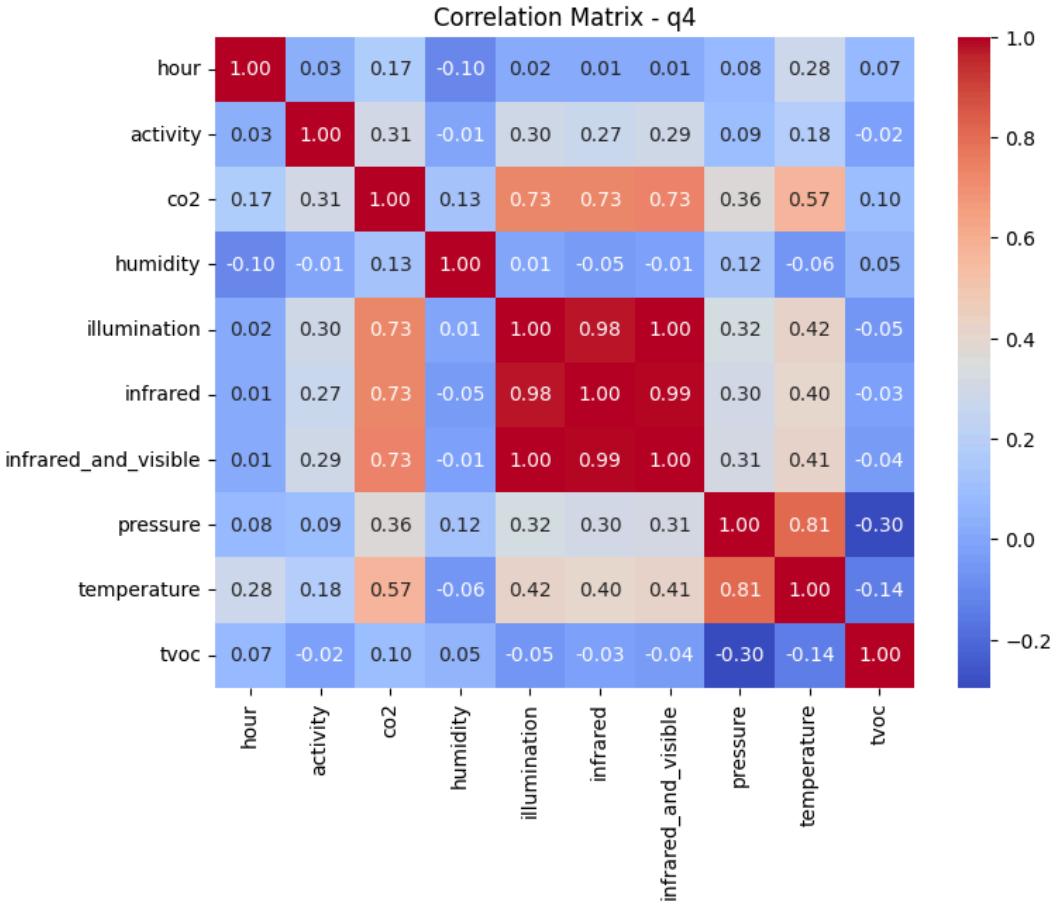


Figure 2: Correlation matrix of the parameters captured by the sensor q4-1003-7456. The matrix has the same parameters as rows and as columns and a single value showing the correlation between the parameter on its row and the parameter on its column. The correlation range is [-1, 1].

The analysis of the correlation brought several matrices as the one in *Figure 2*. CO_2, illumination, infrared, and infrared and visible light tend to increase significantly as more people occupy the room. This correlation is logical, as higher CO_2 levels and increased light and infrared emissions are typically associated with a greater number of individuals present. Additionally, temperature shows a moderate positive correlation with these factors, likely due to the increased thermal output generated by more occupants in the room.

Moreover, there's a weak positive correlation between activity levels and CO_2, illumination, infrared, and infrared and visible light. This suggests that when classroom engagement is higher, these environmental factors experience a slight uptick.

Conversely, humidity exhibits a weak negative correlation with CO_2, illumination, infrared, and infrared and visible light. This phenomenon could be attributed to the moisture expelled by individuals, causing a minor decrease in these factors as humidity levels rise.

4. Energy consumption & occupation relationship:

In order to look for the relationship between energy consumption and faculty occupation, both datasets were merged together. Data gaps in the occupation dataset (night hours, weekends…) were filled with 0.

Upon initial examination of the raw hourly and daily correlations, both metrics provided values below 0.4, prompting a reevaluation of the analytical approach.

University occupancy fluctuates seasonally, with periods like January and June typically witnessing reduced faculty presence due to holidays and examinations. Notably, summer months exhibit 0 occupation values throughout all days, leading to a decline in the correlation coefficient.

Afterward, the correlation was computed for each month, considering both hourly and daily formats.

	CORRELATION (daily)	**CORRELATION (hourly)**
January	0.26	0.39
February	0.53	0.51
March	0.57	0.55
April	0.51	0.49
May	0.48	0.35
June	0.04	-0.04
July	--	--
August	--	--
September	0.55	0.52
October	0.70	0.79
November	0.75	0.77
December	0.78	0.87

Table 1: Table with 12 rows and 2 columns. The rows represent each month and the 2 rows correspond to the 2 types of correlation between energy consumption and occupation computed: Daily and hourly

The correlation analysis (see *Table 1*) revealed distinct patterns in the relationship between energy consumption and faculty occupation throughout the year, reflecting various external factors influencing campus dynamics.

January and June, characterized by holidays and exam periods, showed relatively lower correlations. These periods likely contained 0 occupation values, even if there are students in the faculty doing exams.

During the summer months, with an occupation value of 0, the absence of faculty further diminishes correlations. Additionally, from May until July, and in September, the air conditioning systems are turned on, significantly altering energy consumption dynamics. Notably, since heating

is facilitated by gas during the rest of the months, it does not contribute significantly to energy consumption variations, explaining the observed patterns.

These conclusions highlighted the relevance of the annotations of faculty occupation taken by the university, which do not handle properly 'special' periods, leading to lower metrics.

All those things can be easily observed in *Figure 3*.

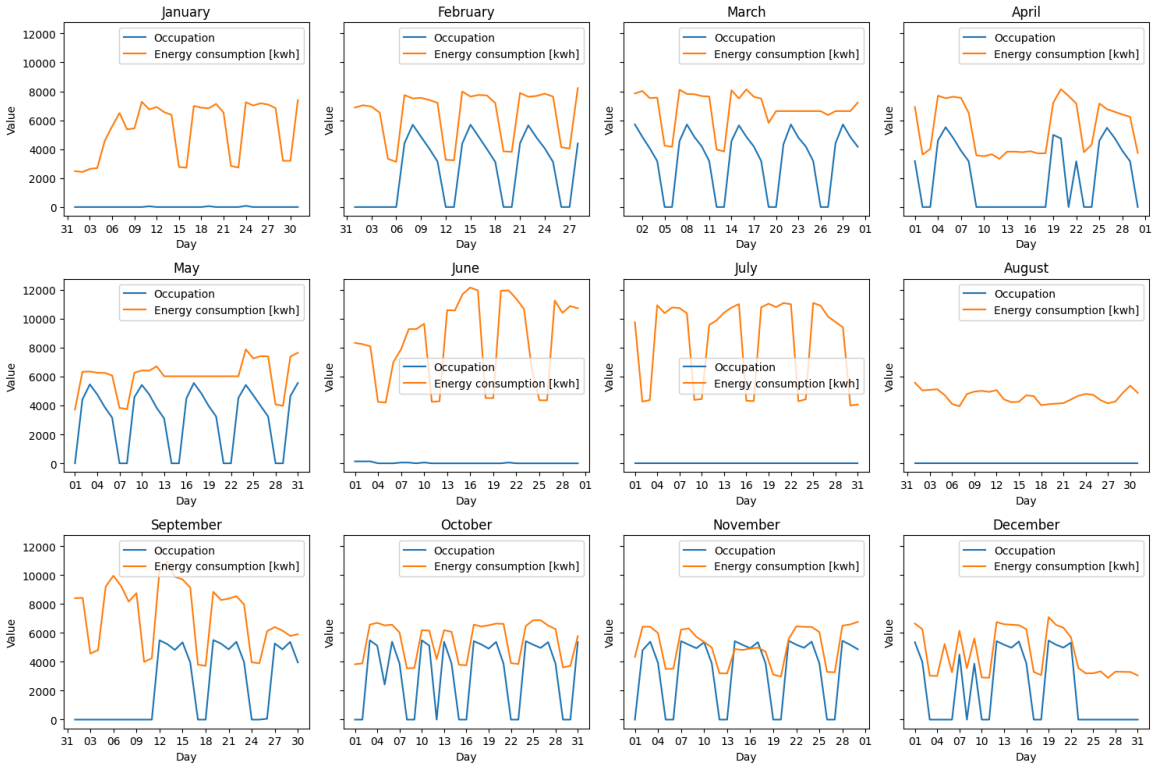


Figure 3: Subplot showing the occupation (blue lines) and the energy consumption (orange lines) across the 12 months of 2022. The x axis of each subplot represents the day of that particular month. The y axis contains the value of that variable in that day.

5. Predictions for occupation and energy consumption:

After analyzing the relationship between energy consumption and faculty occupation throughout the year, the decision on which prediction approach to adopt was taken. Some months showed unexpectedly low values, indicating that energy consumption couldn't reliably predict occupation data, and vice versa.

Several approaches were considered for both energy consumption and faculty occupation data.

Initially, various machine learning algorithms were tested using raw data, attempting to predict energy consumption based on historical records. However, their performance was notably poor due to the dependency of this data on a specific month and day of the week. Considering the very low metrics achieved, a different path was taken to predict energy consumption and occupation.

After that, a seasonal algorithm was considered. SARIMA (seasonal autoregressive integrated moving average) forecasts data by analyzing the patterns and trends in past data and

autocorrelation. The confidence interval (CI) around the forecasted mean represents the range of likely outcomes for energy consumption, given the model's uncertainty. A narrower CI indicates higher confidence in the forecast, while a wider CI suggests greater uncertainty. The forecast provides not only a point estimate but also a range of possible values, helping to assess the reliability of predictions. It accounts for variability in the data and uncertainty in the model's parameters.

Monthly data was collected from January 2018 until half of February 2024, so the model tried to predict the energy consumption of the last 12 months. Due to the complexity of the data patterns of daily (season was 365 days) and hourly data, it only provided slightly accurate predictions for the monthly energy consumption.

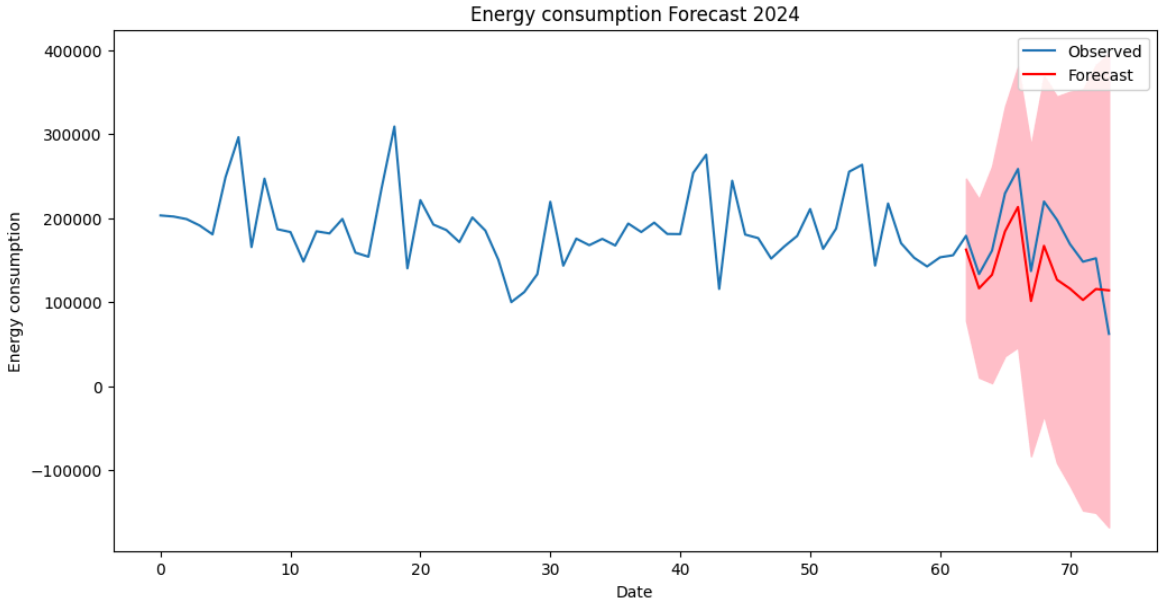


Figure 4: Plot showing SARIMA performance predicting the energy consumption in kWh during the months of 2024. The blue line represents the ground truth while the red line represents the prediction accompanied by a red zone indicating the Confidence Interval.

Subsequently, a rule-based approach was implemented. This algorithm focuses on both the day of the week and the month, adjusting for the lower data values during the pandemic and the subsequent year. This method proved promising predictions.

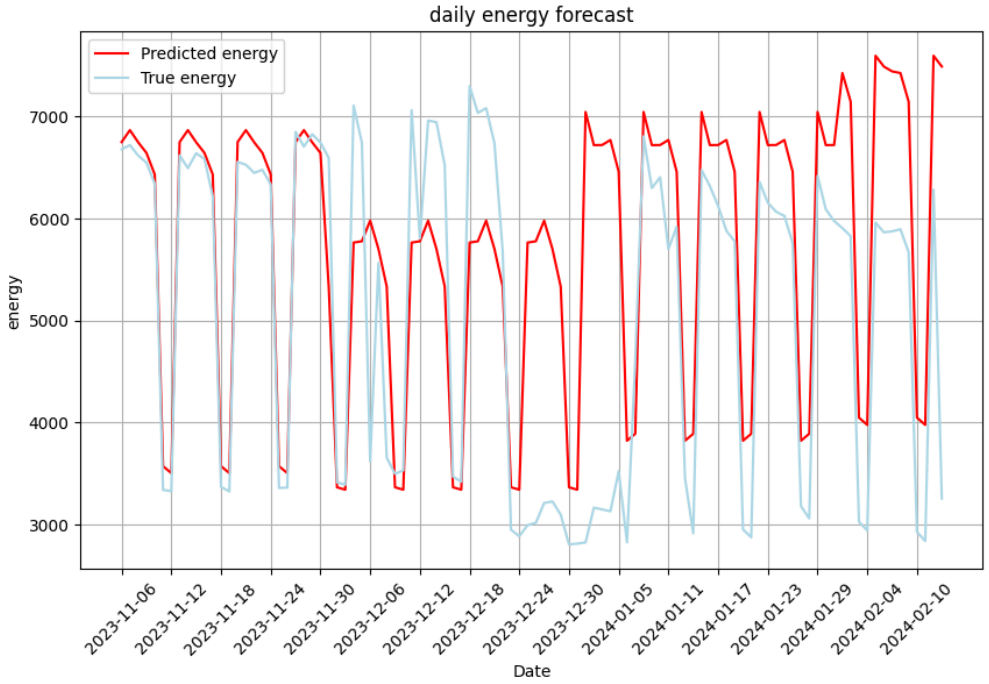


Figure 5: Plot showing both the predicted (from the rule-based approach) (red line) and true values (blue line) of the energy consumption from the final months of 2023 to beginning of 2024.

The difference between the ground truth and the predicted value in *Figure 5* is attributed to the special periods of the year. In the case of December, half of the month is holidays, and the other part is not. Since the rule-based method computed the mean of a specific weekday of the specific month, the predicted value will be lower for working days, and higher for festive ones; which means there's still room for improvement.

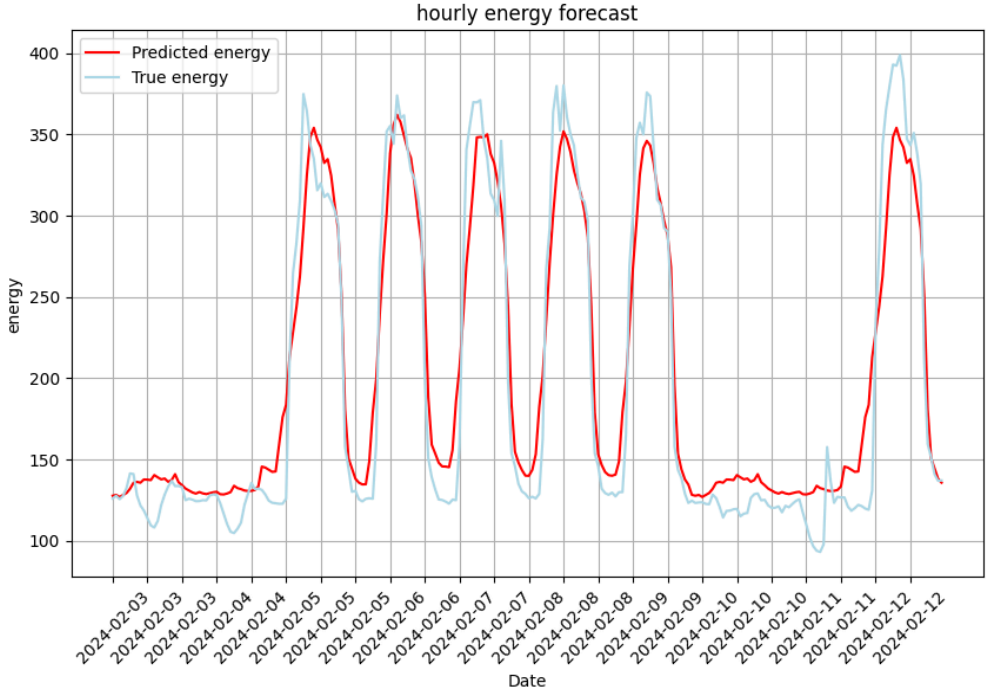


Figure 6: Plot of the hourly energy consumption prediction (red line) and the true values (blue line) from the first days of February 2024 to 10 days later.

As it can be seen in *Figure 6*, the algorithm provides good predictions for this specific period. Moreover, the same process was followed in order to compute the occupation dataset prediction. Finally, MAE, MSE, and RMSE were computed to get quantitative values over the graphs.

Prediction Type	Mode	MAE	MSE	RMSE
Energy Consumption	Daily	932.242	1.787e+06	1337.137
Energy Consumption	Hourly	20.363	7.180e+02	26.797
Occupation	Daily	768.082	1.675e+06	1294.292
Occupation	Hourly	35.972	8.863e+03	94.146

Table 2: Table showing evaluation measures over the predictions computed with the rule-based method. The 3 computed metrics are Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE)

The scale of the data needs to be considered when analyzing the performance of the models; with larger data values, it is normal to obtain larger error metrics (MAE, MSE, RMSE) because these metrics measure absolute errors.

Given that the Rule-based approach worked well when the weekday was specified and after some guidance, the following step was to retry to train machine learning methods this time giving information about the weekday. Three models were tested in order to have a varied range and try to achieve the best performance.

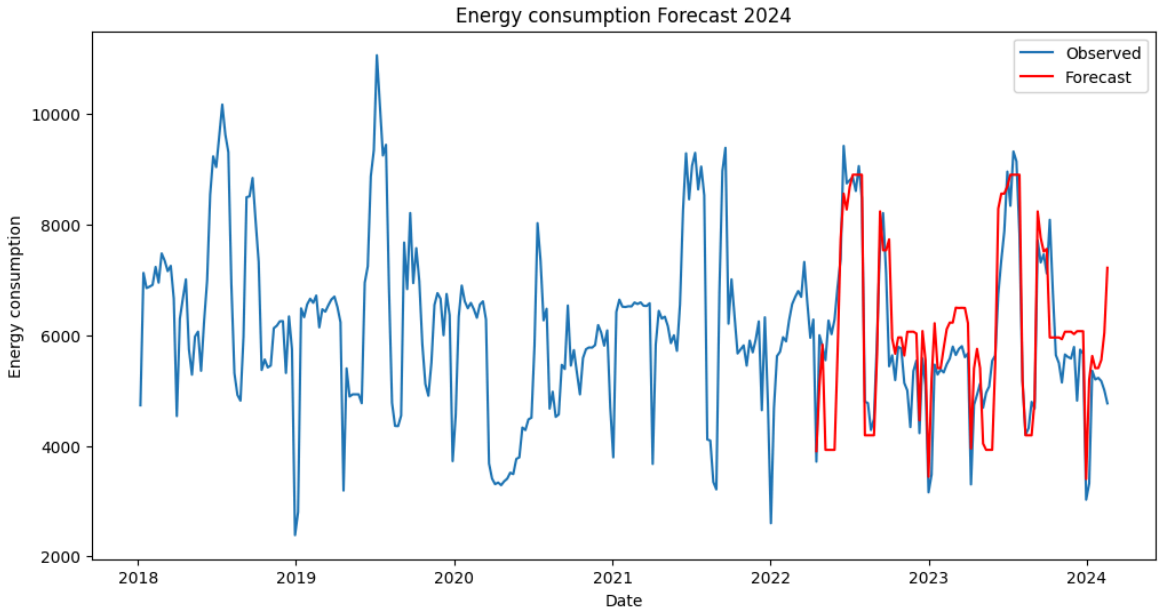


Figure 7: Plot of the energy consumption prediction with Random Forest Regressor model across the entire dataset (except the predicting dates). The blue line represents the ground truth while the red line represents the prediction.

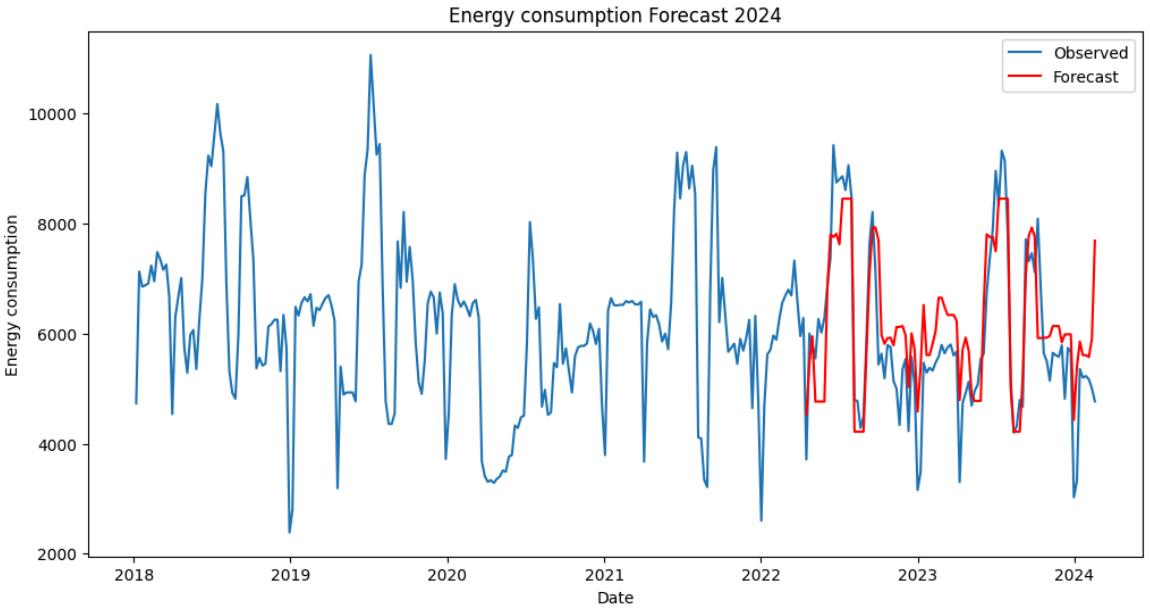


Figure 8: Plot of the energy consumption prediction with Gradient Boost Regressor model across the entire dataset (except the predicting dates). The blue line represents the ground truth while the red line represents the prediction.

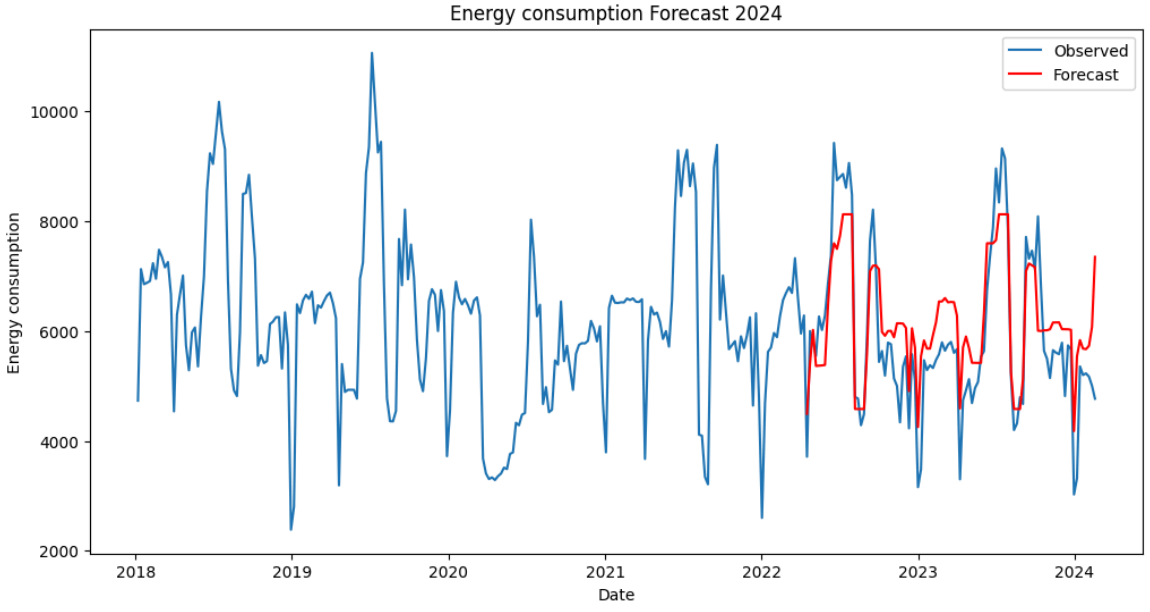


Figure 9: Plot of the energy consumption prediction with Extreme Gradient Boost Regressor model across the entire dataset (except the predicting dates). The blue line represents the ground truth while the red line represents the prediction.

As seen in *Figure 8, Figure 9*, and *Figure 10*, the three models captured the patterns of energy consumption. Nonetheless, the Random Forest Regressor seemed to be the only one that accurately predicted flowing inside the correct range of values. The remaining two models did not correctly get the lower and higher peaks which are crucial when working with energy

consumption. Generally, the user will want to know when the consumption will be the highest or the lowest to take fitting measures according to the foreseen data.

After the past 3 models and the SARIMA, evaluation metrics were computed in order to obtain numeric values to accurately decide which was more fitting.

Model	MAE	MSE	RMSE
SARIMA	41602.139	1.959e+09	44267.820
Random Forest Regressor	873.155	1.443e+06	1201.439
Gradient Boost	951.808	1.502e+06	1225.683
Extreme Gradient Boost	909.749	1.376e+06	1173.328

Table 3: Evaluation metrics of the 4 machine learning models given. The 3 computed metrics are Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE)

6. PV production dataset:

The final dataset essential for the analysis was the PV production dataset. It encompassed comprehensive information about the photovoltaic output of the solar panels installed at the faculty. Initially, the analysis was conducted using data from a 4-month period. However, additional data was subsequently gathered, extending the dataset to encompass up to 8 months of production, from September 2023 until the 28th of April.

The first plots of September, October, November and December demonstrated a notable peak in September compared to subsequent months. This higher production can be attributed to the combined influence of warmer temperatures and longer daylight hours characteristic of the month.

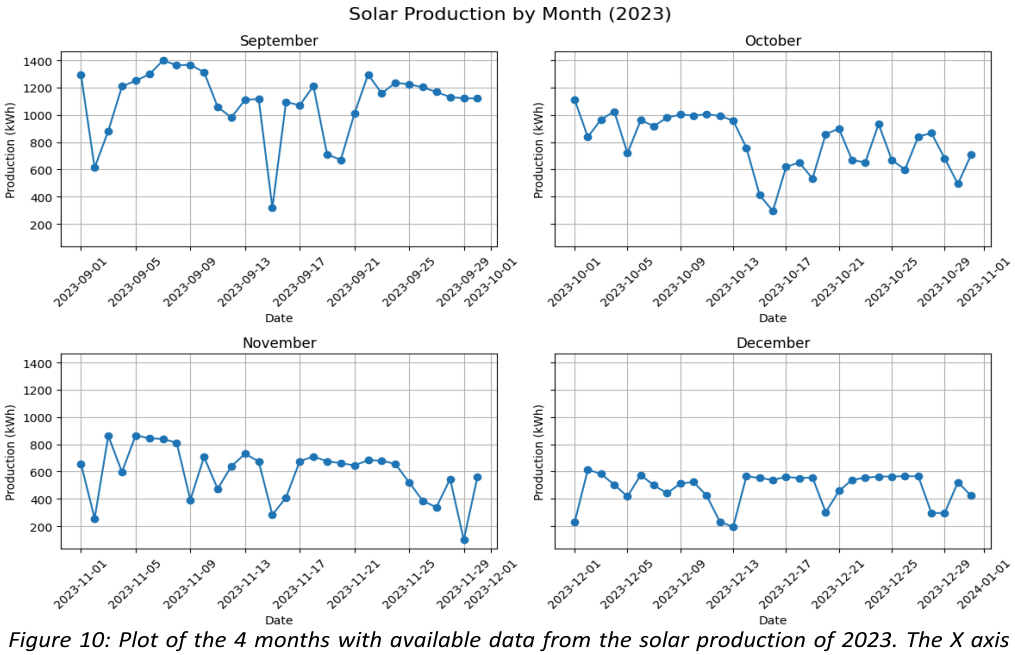


Figure 10: Plot of the 4 months with available data from the solar production of 2023. The X axis represents the date of the month while the y axis shows the production values in kWh format

The plots from January to April indicate that during the winter months, production does not reach high levels, with maximum values approximating 850 kWh in January and 1100 kWh in February. In contrast, production exceeds 1600 kWh in March and reaches up to 2000 kWh in April. This demonstrates a clear seasonal pattern where energy production increases as winter recedes.

Furthermore, a similar pattern is observed within each of these months, where there are days of substantial production and days of significantly lower production. This variability is consistent throughout the winter and early spring months, reflecting fluctuations that are likely influenced by changing weather conditions and daylight availability.

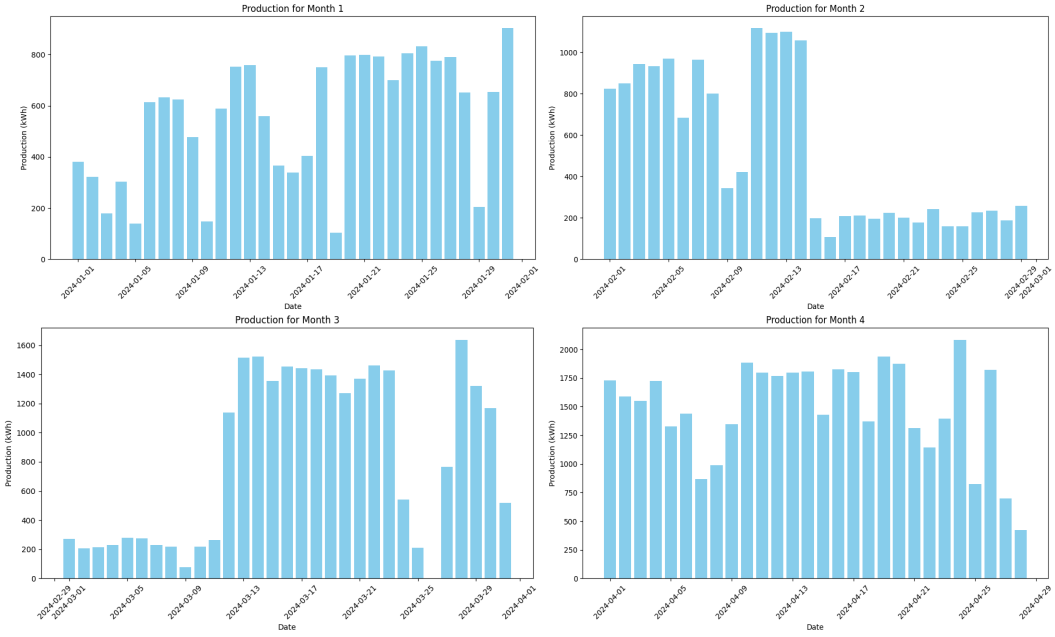


Figure 11: Plot of the next following months with available data from the solar production of January (Month 1), February (Month 2), March (Month 3) and April (Month 4). The X axis represents the date of the month while the y axis shows the production in kWh.

Therefore, the observed lower levels of PV production in some days necessitated further analysis; investigating the correlation between external factors such as outside temperature and weather patterns with PV production. Such analysis could provide insights into the interplay between environmental conditions and renewable energy generation, facilitating more effective planning and optimization strategies.

To understand the reasons behind these fluctuations, it was necessary to include an additional dataset related to weather conditions. This analysis could reveal how factors such as sunlight, temperature, and precipitation cover influence production levels.

The AEMET FABRA dataset, with data from the *Fabra Observatori* in *El Tibidabo* was chosen as it was the nearest location that gave complete and accurate data. This dataset included the following variables for each day:

- fecha: Date
- indicativo: Climatological Indicator
- nombre: Station Location
- provincia: Province
- altitud: Altitude
- tmed: Daily Average Temperature
- prec: Precipitation (mm)
- tmin: Minimum Temperature
- horatmin: Time of Minimum Temperature
- tmax: Maximum Temperature
- horatmax: Time of Maximum Temperature
- dir: Maximum Gust Direction (tens of degrees)
- velmedia: Average Wind Speed (m/s)
- racha: Maximum Wind Gust (m/s)
- horaracha: Time of Maximum Wind Gust
- sol: Insolation (hours)
- presMax: Maximum Pressure (hPa)
- horaPresMax: Time of Maximum Pressure
- presMin: Minimum Pressure (hPa)
- horaPresMin: Time of Minimum Pressure
- hrMedia: Daily Average Relative Humidity (%)

From these variables, the following were selected as the most relevant for the analysis:
- fecha: Date
- tmed: Daily Average Temperature
- prec: Precipitation (mm)
- tmin: Minimum Temperature
- tmax: Maximum Temperature
- sol: Insolation (hours)

These selected variables were crucial for investigating correlations between weather conditions and energy production fluctuations, providing insights into how temperature, precipitation, sunlight, and temperature extremes influence production patterns over time.

The robustness of the methodology was ensured by evaluating the correlation between temperature and PV production. The Pearson correlation coefficient between the mean temperature (tmed) and PV production was found to be 0.4404951206335189, indicating a moderate positive association. Similarly, the correlation between the minimum temperature (tmin) and PV production was 0.3633402405372172, which was slightly weaker but still positive. The maximum temperature (tmax) exhibited the strongest correlation with PV production at

0.49331066617618863. These results suggested that as temperature increased, PV production tended to increase as well, with the maximum temperature having the most significant impact among the three temperature measures.

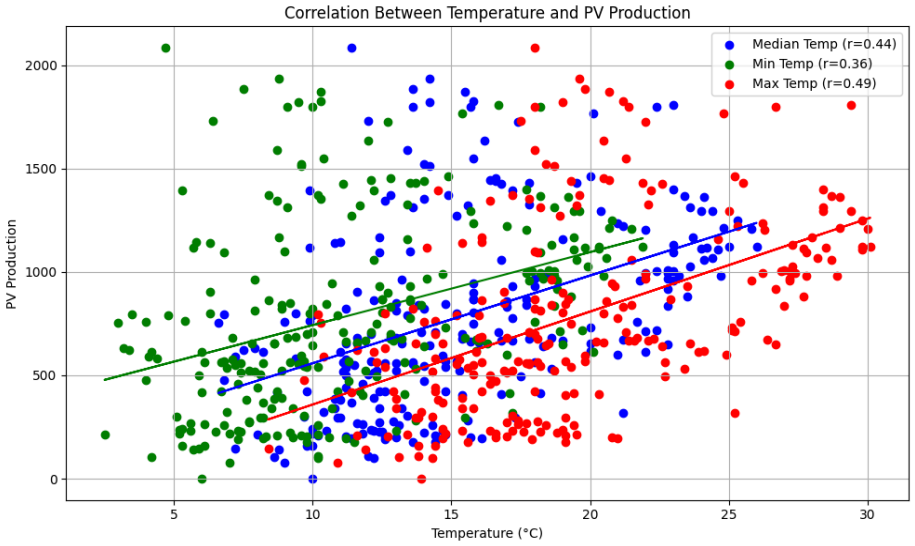


Figure 12: Plot of the correlation between temperature and PV Production. The X axis represents the temperature in ºC while the y axis shows the PV production in kWh.

In the subsequent phase of the analysis, the meteorological dataset was augmented to address missing entries in the "Sol" variable, initially considered to be quite important. To resolve this issue, a machine learning approach was implemented to extrapolate and estimate the missing values based on available data. Various machine learning models were tested to ascertain their efficacy for this task. Specifically, the performance of the following models was assessed:

- K-Nearest Neighbors (KNN)
- Extreme Gradient Boosting (XGB)
- Multilayer Perceptron (MLP)

Model	MAE	MSE	R^2 SCORE
K-Nearest Neighbors	2.119	7.589	0.466
Extreme Gradient Boosting (XGB)	2.022	6.935	0.512
Multilayer Perceptron (MLP)	1.774	5.543	0.610

Table 4: Evaluation metrics of the three different models to fill the Sol variable.

The MLPRegressor model exhibited the most favorable performance metrics, including the lowest Mean Absolute Error, Mean Squared Error, and highest R^2 Score among the models evaluated. This indicated superior accuracy in predicting the missing "Sol" variable values.

With the interpolated values a final list was created with the filtered data for months 9,10,11,12 of 2023 and 1,2,3,4 (until the 28th of April) of 2024.

7. PV predictions:

With the missing values successfully extrapolated, a comprehensive JSON file was created containing all the necessary data for an 8-month period. This dataset included complete records of temperature, precipitation, "Sol" values, and PV (photovoltaic) production.

Building on this foundation, the next step was to train a model to predict PV production. To achieve this, meteorological data from the years 2019 to 2024 was utilized. By leveraging this historical weather data, the model was trained to make accurate predictions of PV production based on meteorological conditions.

To initiate the training phase, several models were tested and evaluated to identify the one with the best performance. The first assessed model was the Random Forest Regressor. Random forest was chosen mainly for two reasons: it can capture nonlinear complex relationships unlike linear models while not being overly complex, and it is easily scalable to larger datasets, which is beneficial if the model is used with more data in the future. Deep learning approaches were not tried due to the limited amount of data available. After training the initial random forest model, the following results were obtained, as illustrated in the accompanying plot:

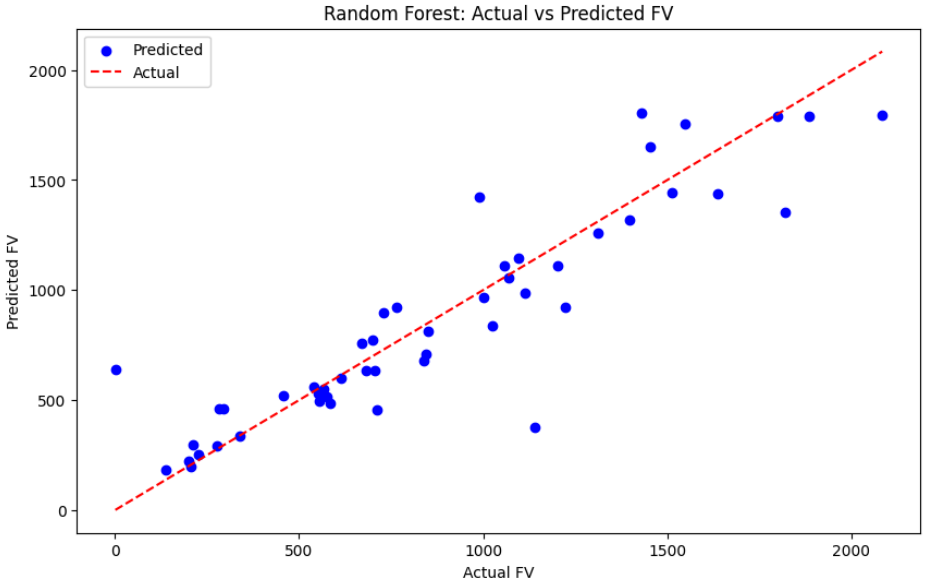


Figure 13: Plot of the comparison between the actual and predicted FV with the Random Forest Regressor model. The X axis represents the actual PV production while the y axis shows the predicted PV production in kWh.

From the last figure, it is observable that the Random Forest Regressor model exhibited better performance when actual photovoltaic (PV) production values were lower. This trend likely arises from the absence of data from summer months in the dataset, where PV production tends to be higher. Consequently, as PV values increased, the model's predictions deviated significantly from the red dashed line in the latter part of the graph. The results of the random forest model were satisfactory; nevertheless, other models were implemented and their RMSE calculated for comparison:

Model	RMSE
Gradient Boosting	36126.121
Extreme Gradient Boosting (XGB)	39834.058
Multilayer Perceptron (MLP)	57324.558
SVR	55801.698

Table 5: Evaluation metric (RMSE) of the different models

The model with the lowest RMSE, and therefore the best performance, is the Gradient Boosting model. This model was chosen to compare with the Random Forest; however, it did not improve the overall performance. The following plot illustrates the performance of the Gradient Regressor:

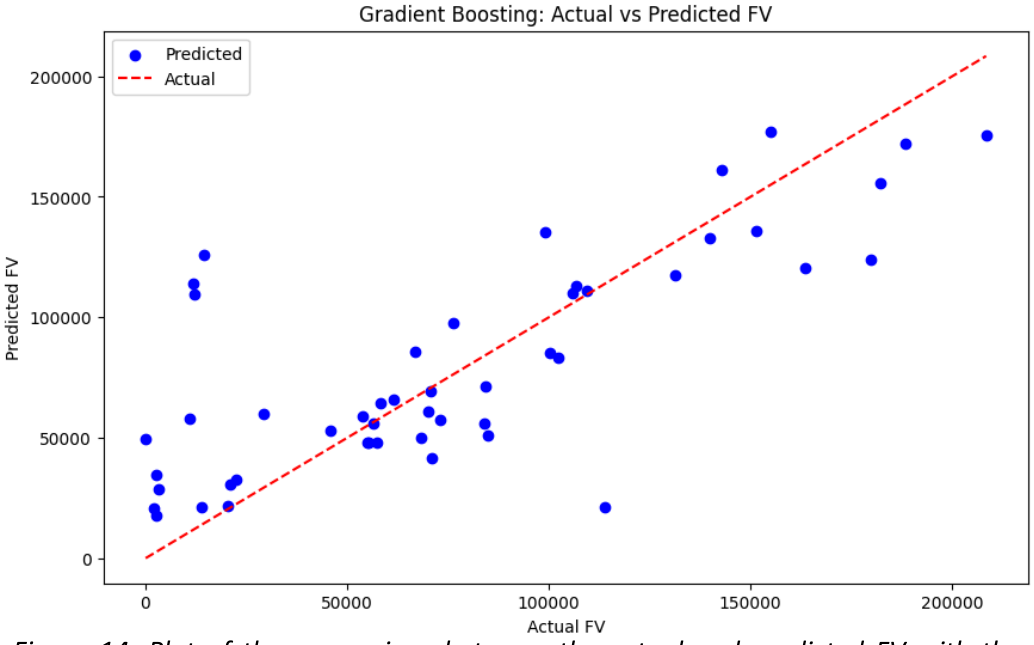


Figure 14: Plot of the comparison between the actual and predicted FV with the Gradient Boosting model. The X axis represents the actual PV production while the y axis shows the predicted PV production in kWh.

To improve predictions, and given that using different models did not help, predictions were attempted without the "Sol" value. This decision was based on the assumption that the "Sol" value might introduce errors and noise, impacting the accuracy of predictions due to accumulated error. A Random Forest Regressor was trained and evaluated without including the "Sol" variable, following the same procedure as before. This model was ultimately chosen because, if scaled in the future, the predicted "Sol" variable could introduce complex errors.

The model was then implemented into the project's website, enabling daily updates with new weather data. An XML file provided daily by AEMET was identified as a solution for this requirement. It can be found at the following link:

https://www.aemet.es/xml/municipios/localidad_08266.xml

The problem with this file is that the data included is from Cerdanyola, and not from *Observatori Fabra*, where our training data came from. Nevertheless, we need to do this trade off, because data from *Observatori Fabra* had a complete historical record of data with all the variable we needed, but it is not everyday actualized with weather predictions, whether the data from Cerdanyola is actualized everyday with predictions but we didn't find a historical archive with complete weather data with the variables we needed. The XML data was parsed, and relevant meteorological features were extracted and converted into JSON format. This data was then used to make predictions with the pre-trained model.

A function was defined to preprocess new input data by converting dates to ordinal and ensuring all necessary features were present. Another function loaded the trained model (best_model.pkl) to make predictions based on the preprocessed data. The results included predicted FV values for the given dates.

A comprehensive metrics analysis was conducted. The obtained results were satisfactory:

- Number of samples in the training set: 192
- Number of samples in the test set: 49
- Number of features: 6

Performance Metrics:

- Mean Squared Error (MSE): 44800.968
- Root Mean Squared Error (RMSE): 211.662
- Mean Absolute Error (MAE): 139.104
- R-Squared (R^2): 0.827
- Explained Variance Score: 0.828

Actual vs Predicted 'fv' values (first 10 samples):

Date	Actual FV	Predicted FV
2023-09-25	1223.30	922.80
2023-09-07	1398.08	1318.45
2024-04-10	1884.73	1792.15
2024-03-27	763.95	921.65
2024-04-24	2083.94	1796.96
2024-02-21	201.60	223.84
2024-03-16	1454.80	1653.44
2024-04-08	991.01	1422.15
2023-09-10	1312.98	1260.00
2023-12-23	554.84	494.71

Figure 15: Representative sample of the obtained results for the comparison between predicted and actual FV values.

It can be concluded that the model explained a significant portion of the variance in the target variable (FV), indicating strong performance. The absence of the "Sol" variable did not significantly degrade model performance, leading to the decision to eliminate it. Additionally, the “fecha” variable and temperature-related features were identified as critical for accurate predictions.

Recommender system:

The recommender system developed for the Q4-1003 class utilized sensor data to offer personalized insights and warnings tailored to the environment's conditions. Leveraging a combination of sensor variables including CO_2 levels, activity, humidity, illumination, infrared, pressure, temperature, and TVOC, the system employed simple algorithms to determine actionable recommendations.

These recommendations were contingent upon factors such as the current season, ensuring that interventions like opening windows to alleviate high CO_2 levels or TVOC were contextually appropriate. Additionally, the system accounted for university operating hours, advising users when no actions were needed during closed periods. By incorporating insights on natural lighting availability and adjusting recommendations accordingly, the system aimed to optimize environmental conditions for comfort and well-being. Implemented as part of a web application, the system delivered real-time warnings to users, empowering them to make informed decisions.

Web:

A webpage was developed to display the collected data and enable energy managers to make future predictions based on historical trends. The webpage is divided into four tabs:

- Sensors: this section allows users to select from one of the seven sensors placed throughout the university campus. Each sensor has a display of variables and interactive plots that show the collected data, including real-time information. This feature provides a visual understanding of the patterns and an interface for users to monitor real-time metrics.

- Predictions: this section is designed to enable users to predict future data. It supports various prediction methods, including energy consumption and faculty occupation, and allows users to choose between daily or hourly modes and set the end date for the prediction. After the user submits the form, the code is executed, and the results are displayed. The outcomes include a plot illustrating the predictions, a displayed data frame, and the option to download a CSV file for easier usage.

- PV predictions: separate from the former tab, the PV predictions tab is designed with a machine learning model to give PV production predictions for the following 6 days based on weather predictions from the AEMET official site from Cerdanyola. The tab has a table with the predicted weather data including mean temperature, max temperature, min temperature, and precipitation predictions. Additionally, it displays 2 plots: the first one merging the weather predictions and the second displaying the PV production predictions.

- Recommendations: it displays metrics for the Q4-1003 classroom, where the predictions are made. It provides possible recommendations based on these metrics, helping users to make informed decisions.

The website uses an “app.py” main file that renders information to the correct tab accordingly. It is the center of operations. The performing tasks programmed include:

- Obtention and storage of the sensor data to a SQLite table
- Creation of the plots for all the different parameters of each sensor
- Prediction of both the occupation and energy consumption values for hourly mode or daily mode
- Prediction of PV production
- List creation of recommendations.

It is worth mentioning that for the prediction of energy consumption and occupation, the website uses the rule-based method. Given that both the rule-based method and the random forest regressor took more or less the same time on the trials to predict the values, the rule-based method was the one used due to having easier transportation to the website environment. The code was copied and pasted from the working notebook to the website instead of exporting and importing the model which would mean more files to add to the website environment.

Additionally, other files contributed to the creation of the website. An HTML file for each section of the web was created to arrange the display of information. This file worked in tandem with a CSS file that contained the design of all the elements of the website.

The website works with libraries like the previously mentioned SQLite as well as Flask which allows for easy management of the website data. Once programmed, the website was published to the Internet via Oracle and can be found at http://143.47.57.135:5000/.

Conclusions and open challenges:

The project successfully achieved its primary objectives, demonstrating the feasibility and effectiveness of using sensor data and predictive algorithms to manage energy consumption at the university. Key accomplishments include:

1. Data integration and analysis: various datasets, including energy consumption, faculty occupation, and sensor data, were successfully integrated to analyze the patterns and correlations. This integration provided a comprehensive understanding of the energy dynamics within the university buildings.

2. Predictive modeling: several predictive models were evaluated to forecast future energy consumption and faculty occupation. After some refinement, a rule-based approach and various machine learning models considering the day of the week and month were subsequently implemented, yielding accurate predictions.

3. Real-time monitoring and forecasting: a user-friendly web application was developed to display real-time data and predictions, making it accessible for energy managers to monitor and forecast energy consumption. The application features interactive plots, sensor metrics, and prediction tools, enhancing decision-making capabilities.

4. Recommender system: a recommender system was also developed for optimizing classroom conditions based on sensor data. This system provides actionable recommendations, such as adjusting ventilation or lighting, to improve environmental comfort and energy efficiency.

Despite these successes, several challenges and limitations were encountered, presenting opportunities for future work:

1. Real-time data integration: the project was limited by the lack of real-time energy consumption and PV production data. Future efforts should focus on integrating these real-time data streams to provide more accurate and timely recommendations for energy savings.

2. Data cohesion: The team had to deal with various data types and formats, which required a significant amount of data preprocessing and cleaning.

3. Website creation: The development of the website was another major challenge, as there was no prior experience in this area. However, through an exterior party and self-learning, a user-friendly website was created and uploaded to the Internet.

4. Model accuracy and improvement: while the rule-based approach showed promise, there is still room for improvement. Special periods, such as holidays and exam times, significantly impacted prediction accuracy. Developing more sophisticated models that can account for these anomalies will enhance prediction reliability.

5. Scalability and generalization: the models and systems developed were tailored specifically for the university's context. Extending the approach to other buildings or institutions requires further validation and potentially different configurations to account for varying usage patterns and building characteristics.

6. User engagement and feedback: ensuring consistent use of the web application and implementation of recommendations by energy managers and other stakeholders is crucial. Future work could involve user training and the incorporation of feedback mechanisms to refine the system based on actual user experiences and needs.

7. Limited Historical Solar Panel Production Data: Another significant challenge encountered was the lack of comprehensive data on the production of solar panels, with only 8 months of production records available. Notably, these missing months predominantly encompassed periods of higher production, such as the summer months. This limitation posed challenges in accurately predicting solar panel output during peak periods. However, it is anticipated that this issue will be resolved in the future as more comprehensive data becomes available.

Addressing these challenges can significantly enhance the project's impact, contributing to more sustainable and efficient energy management practices within the university and beyond.