**Executive Summary: IE Sustainability Datathon 2025**

**Team: CTRL + ALT + DEFEAT**

**Members:** Joy Zhong, Joel James Alarde, Maria do Carmo Brito e Abreu, Emiliano Puertas Ajuria, Uxía Lojo Miranda, Sebastián Felipe Zambrano Julio

**1. Problem Overview & Business Context**

Spain’s energy transition highlights a growing gap between renewable energy production and real-time consumption. With the Spanish government aiming for 160 GW of installed renewable capacity by 2030—138 GW of which will be wind and solar—balancing production with demand has become a national priority. However, the intermittent nature of renewable sources, such as solar and wind, creates challenges in synchronizing generation with consumption. In 2022 alone, 3.3 TWh of wind energy was wasted in Spain because it could not be integrated into the grid at the time it was generated. This scenario underscores the urgent need for flexible solutions like energy storage and intelligent demand management.

At the core of this challenge lies the Repsol industrial facility under analysis, which has a 175 kWp photovoltaic installation dedicated solely to self-consumption. The installation is completely off-grid, meaning any energy generated that is not immediately consumed is permanently lost. This limitation results in the underutilization of clean energy that could otherwise displace more carbon-intensive electricity from the grid.

Having said that, our project supports Repsol’s goal of minimizing renewable energy waste and driving its decarbonization agenda through a structured, data-driven approach. By forecasting energy potential, intelligently scheduling storage and discharge cycles, and evaluating emissions offsets, we translate raw industrial data into actionable sustainability outcomes.

**2. Project Objectives**

* **Objective 1:** Predict the maximum hourly solar generation potential for September 2024 using weather data and calculate the underutilized solar energy.
* **Objective 2:** Simulate the usage of a theoretical 100 kWh battery to optimize solar energy storage and calculate the Self-Consumption Ratio (Ra).
* **Objective 3:** Calculate the total CO2 emissions avoided (CO2ev) by using the battery to replace grid electricity during high-emission periods.

**3. Data Overview**

Our project integrated a total of six key datasets, each of which contributed unique information essential for modeling solar energy generation, consumption, and environmental impact. Through this robust data preparation pipeline, we ensured each dataset was fully optimized for its corresponding objective: solar forecasting (Obj. 1), energy recovery simulation (Obj. 2), and environmental analysis (Obj. 3). For **Objective 1**, we used Generacion\_fotovoltaica.csv, which includes 10,320 hourly records of actual solar generation from July 24, 2023 to August 31, 2024, and Meteorologia.xlsx, a weather forecast dataset containing 27 meteorological variables across four latitude-longitude points near the factory. For **Objective 2**, we incorporated Consumo\_fotovoltaica.csv, which contains 720 hourly values (September 2024) of how much solar energy the factory actually consumed. For **Objective 3**, we analyzed Consumo.csv (11,040 hourly entries of grid energy use) and carbon intensity files ES\_2023\_hourly.csv and ES\_2024\_hourly.csv, with 8,760 and 8,784 hourly rows respectively, representing Spain’s carbon intensity data over 2023 and 2024.

**4. Exploratory Data Analysis**

Preprocessing and the EDA phase for our data was critical in order to ensure compatibility across datasets. To start, all timestamps were standardized to a single format (FORECAST\_TIMESTAMP) and converted to pandas datetime objects in UTC+1 (Madrid time). We removed timezone indicators in order to obtain a uniform index for our final dataframe. Dataframes were then merged on a common hourly time index using a left outer join.

Once the data frames were merged, we checked for duplicate indexes and noticed we had many - 89,384. In order to take care of these duplicate values, we decided to go for an aggregation approach. The first thing we did was we obtained the latitude and longitude value of the city Pinto, which is where the Factors is located, and then we found the 4 latitude-longitude pairs in the dataset that were the closest to the values associated with Pinto. We kept all rows that had those 4 unique pairs and then we dropped the rest, since they wouldn’t be relevant for our forecast. Then for the remaining duplicate indexes, we averaged all the feature values to keep only one of the timestamps in order to get rid of duplicate indexes. There were no null feature values after this data cleaning. For the null values of our target, TOTAL\_KWH\_ENERGIA, we used a model-based imputation, which we will explain in detail further below in this report.

We implemented a Smart Outlier Strategy grounded in domain knowledge and data diagnostics. Outlier detection was conducted using the **Interquartile Range (IQR) method**, which flagged values falling significantly outside the expected range. Rather than removing these outliers outright, we used this analysis to understand the variability of each feature and retained values unless they were physically implausible. For example, in the energy datasets, small negative values between -1 and 0 kWh were preserved, interpreted as minor parasitic loads required to keep the photovoltaic system operational during non-generative periods (e.g., at night). In order to better understand our data, we generated a correlation matrix to see how our features correlated with our target - TOTAL\_KWH\_ENERGIA. During our outlier detection, we noticed tpsurface\_0 had a very high percentage of outliers (>20%) and we can also see in the matrix that it is very lowly correlated. So with both of that information, we decided to drop this feature.

**5. Feature Engineering & Feature Selection**

We also engineered several important features:

* **Datetime-derived features:** hour of day, day of week (used for analyzing consumption cycles)
* **Solar excess:** calculated as SOLAR\_KWH\_GENERATED - SOLAR\_KWH\_CONSUMED
* **Battery energy metrics:** columns, such as ENERGY\_CHARGED\_KWH, ENERGY\_DISCHARGED\_KWH, and EXCESS\_ENERGY\_KWH
* **CO₂-adjusted metrics:** used hourly carbon intensity to compute CO2\_EMITTED and CO2\_AVOIDED

This structured preprocessing ensured that each dataset was aligned to support its respective objective effectively.

**4. Smart Outlier Strategy and Feature Selection**

Our preprocessing pipeline began with a careful feature selection process to ensure that only relevant and non-redundant variables were fed into the model. We first conducted a correlation analysis to evaluate how strongly each meteorological feature was associated with solar generation (TOTAL\_KWH\_ENERGIA). This allowed us to prioritize features with the highest predictive value. To further refine the selection and avoid multicollinearity, we applied the **Variance Inflation Factor (VIF)**. Features with high VIF scores—indicating significant overlap with other predictors—were excluded to reduce redundancy and improve model stability. This two-step selection process ensured that the final feature set was both informative and statistically robust.

To prepare clean and reliable inputs for modeling, we first analyzed the features by grouping them into their respective distances (eg. on the surface, 2m above the surface) and checked their correlation to solar generation. From this, we gathered that the most important features were total\_radiation, SUNSDsurface\_0, and avg\_temp and the least correlated features were dayofweek\_cos, avg\_wind\_u, and gustsurface\_0. Multicollinearity was also handled to see which features are interacting with each other to make the model more direct. Then we dropped the least correlated features as the final step before initializing the predictive model.

**5. Modeling Approach – Objective 1**

To address Objective 1, our team developed a predictive model to estimate the hourly maximum solar generation potential for the month of September 2024. The goal was to determine how much energy could have been generated under optimal weather conditions, and compare this potential with the actual solar energy consumed by the factory during the same period.

We tested multiple regression models including XGBoost Regressor, Random Forest, LightGBM, and Stacked Regressors. Final model selection was based on cross-validation results, with a stacked model that combined the strengths of Random Forest, SVR, Ridge Regression, and Gradient Boosting yielding the most consistent performance.

Feature selection was guided by meteorological relevance and multicollinearity analysis. Variables such as solar radiation (SUNSDsurface\_0), cloud cover (tccatmosphere\_0), air temperature (2theightAboveGround\_2), wind vectors, and humidity were included. Recursive Feature Elimination (RFE) and Variance Inflation Factor (VIF) filtering helped ensure a robust, interpretable feature set.

The training and test sets were split using an 80/20 partition, and the final stacked model achieved a Mean Absolute Error (MAE) of 5.891 from the training data, representing a refined improvement over earlier baselines. This indicates strong predictive performance given the hourly granularity and weather variability.

We then applied the model to generate predictions for each hour in September 2024, producing a dataset with 720 entries consisting of Datetime and predicted KWH\_ENERGIA. This formed the foundation for underutilization analysis. We compared the predicted solar generation with actual usage from Consumo\_fotovoltaica.csv, resulting in a total underutilization value of 1,142.9573 kWh, representing clean energy that could have been used but was not.

Having said that, these two deliverables, being the predicted\_solar\_generation.csv file and the underutilization metric, constitute the final outputs for Objective 1. They lay the foundation for further optimization through battery simulation and CO2 impact assessment in subsequent objectives.

**6. Battery Optimization – Objective 2**

To address the underutilization of solar energy, we simulated the implementation of a theoretical 100 kWh battery system with a maximum charge/discharge rate of 100 kW, constrained to one full charge and discharge cycle per day. The battery was configured to charge exclusively from solar excess—that is, the hourly difference between predicted solar generation and the factory’s actual solar consumption—and to discharge only to meet internal factory demand. Importantly, the battery remained disconnected from the grid, meaning stored energy could not be exported and was instead strategically deployed to displace grid electricity during high-demand periods.

Our simulation followed a two-step daily cycle. Each day, the battery charged during the earliest hours where excess solar energy was available, until reaching its 100 kWh limit. Once fully charged, the battery discharged later in the day during the hour of highest grid consumption, maximizing its environmental benefit. Through this logic, the battery successfully utilized **1,638.6983 kWh** of solar energy that would otherwise have been wasted. This stored energy was redirected to power the factory, increasing overall solar utilization. As a result, the Self-Consumption Ratio (Ra) improved dramatically—from a baseline of approximately 60% (using only direct solar consumption) to **90.8696%**, representing a more than 30 percentage point increase. This outcome demonstrates the power of intelligent, constraint-aware battery scheduling to significantly enhance renewable energy use and reduce grid dependence in industrial settings.

**7. Environmental Impact – Objective 3**

The third and final objective of this project aimed to quantify the environmental benefits of optimized solar energy use by measuring the total CO₂ emissions avoided. This objective directly connects the technical performance of the battery system with Repsol’s broader sustainability goals, translating self-consumption efficiency into measurable climate impact. The project’s logic was designed to simulate baseline CO₂ emissions and compare them to an optimized scenario in which a 100 kWh battery offsets grid consumption during periods of high carbon intensity. We used hourly carbon intensity data for Spain from the dataset ES\_2024\_hourly.csv, merged with our modeled energy flows from Objective 2.

We began by importing and cleaning the carbon intensity data from ES\_2024\_hourly.csv, which contains hourly records of Spain’s electricity carbon intensity in grams of CO₂ per kilowatt-hour (gCO₂eq/kWh). After merging this data with our master DataFrame containing hourly grid consumption, solar usage, and battery activity, we computed the baseline CO₂ emissions using the formula:

To compute the **baseline emissions**, we multiplied the total hourly grid energy used by the factory in September 2024 by the corresponding hourly carbon intensity values (in gCO₂eq/kWh). This represented emissions under current conditions, without any solar or battery optimization. The total baseline CO₂ emissions for the month were calculated as **11,645,944.24 grams of CO₂**.

To compute the **optimized emissions**, we simulated the battery discharging stored solar energy during high carbon intensity periods, thus offsetting grid usage. The adjusted grid energy (grid energy minus battery discharge per hour) was again multiplied by carbon intensity. The result was an optimized emission profile totaling **11,483,458.13 grams of CO₂**for the same period.

The final metric, **CO₂ avoided (CO2ev)**, was the difference between these two scenarios: **CO₂ avoided = 11,645,944.24 - 11,483,458.13 = 162,486.11 grams of CO₂**. This result demonstrates that even a relatively simple battery charging rule, being focused on high-emission hours, can yield meaningful reductions in greenhouse gas emissions. It validates the feasibility and environmental value of industrial-scale energy storage systems when paired with accurate forecasting and solar excess management. This quantitative output helps transform technical simulation into actionable sustainability insights, aligning with Repsol’s goals of carbon mitigation and renewable energy leadership.

**8. Results & Performance Metrics**

Our modeling and simulation approach delivered strong technical and sustainability outcomes across all three objectives. For solar generation forecasting (Objective 1), our final stacked model achieved a Mean Absolute Error (MAE) of **10.2**, ensuring accurate hour-by-hour predictions of solar potential for September 2024. Based on this forecast, we calculated an **underutilization or wastage of 1,142.9573 kWh**—representing clean solar energy that was neither consumed nor stored by the factory due to limitations in real-time demand and system design.

To mitigate this inefficiency, we simulated the daily use of a theoretical battery system. The battery optimized solar consumption with a total of **1,638.6983 kWh** of additional energy utilized through a smart one-cycle-per-day charge and discharge strategy, directly contributing to factory operations. As a result, the **Self-Consumption Ratio (Ra)**—which measures the percentage of generated solar energy actually used—jumped from a baseline of approximately 60% to **90.8696%**, indicating a substantial gain in energy efficiency.

Finally, the optimized use of stored solar energy allowed us to displace a meaningful portion of grid electricity during peak carbon intensity hours. This strategy resulted in a total of **162,486.11 grams of CO₂ emissions avoided** over the month of September. These results demonstrate the value of combining predictive modeling with smart battery scheduling to improve renewable energy utilization and reduce environmental impact in industrial energy systems.

**9. Key Insights & Strategic Value**

Our analysis provides a strategic framework for enhancing Repsol’s operational efficiency and sustainability through optimized renewable energy storage and usage. The proposed smart battery system delivers benefits across business performance, environmental impact, and societal value.

A strategic battery implementation can recover nearly 2.6 MWh of clean energy lost due to storage limitations, increasing solar utilization by nearly 20% and reducing reliance on high-carbon grid electricity. A 100 kWh battery with a 100 kW charge/discharge rate is technically viable, requiring minimal infrastructure upgrades and ensuring reliable performance with a forecasting accuracy of 6.59 kWh.

Strategic battery discharge during high-carbon grid periods maximizes CO₂ reduction, cutting emissions by over 19 kg of CO₂ monthly and lowering energy procurement costs, particularly during peak pricing.

The scalable, modular design enables expansion across Repsol’s network, enhancing operational efficiency and supporting broader renewable energy goals. Improved self-consumption strengthens ESG performance, increases investor confidence, and reinforces Repsol’s market leadership in renewable energy. Future AI-driven optimization and participation in grid balancing and virtual power plants could create new revenue streams and competitive advantages, positioning Repsol for long-term business resilience and growth.

**10. Limitations & Future Improvements**

While our solution showcases strong technical accuracy and substantial sustainability gains, a few key limitations remain that could be addressed in future iterations. First, our solar generation forecasting is based on **historical weather forecasts**, which are assumed to be accurate. In real-world scenarios, discrepancies between forecasted and actual meteorological conditions could reduce the reliability of solar generation predictions. Incorporating real-time weather updates or nowcasting techniques could improve adaptability.

Second, the battery simulation follows a **rule-based approach**, constrained to one charge and one discharge cycle per day. Although this logic reflects the challenge requirements, it does not dynamically optimize for fluctuating factory demand or real-time carbon intensity trends. In future work, more advanced strategies such as **reinforcement learning**, **greedy algorithms**, or **multi-objective optimization** could be applied to adaptively schedule charging and discharging for greater energy and emissions efficiency.

Third, our framework currently **optimizes for energy and environmental impact alone**, without factoring in economic considerations like electricity prices, battery degradation, or grid tariffs. These elements are essential for assessing financial viability in a real deployment scenario. Additionally, while our ensemble modeling approach (including XGBoost, LightGBM, and a stacked model) yielded strong performance, future improvements could include **time-series architectures** such as LSTMs or temporal convolutional networks to better model the sequential nature of solar generation and grid demand.

Together, these enhancements would help evolve our solution from a robust simulation to a scalable, real-world-ready system for industrial renewable energy optimization.

**11. Conclusion**

This project demonstrates the transformative potential of combining predictive modeling, intelligent storage simulation, and environmental analytics to enhance industrial energy systems. Our end-to-end solution forecasted solar generation with strong precision, identified and quantified underutilized clean energy, and deployed a data-driven battery simulation strategy that captured lost potential and realigned energy flows toward sustainability.

By simulating a 100 kWh battery with realistic constraints, we recovered more than 1.6 MWh of solar energy, boosting the factory’s self-consumption ratio from approximately 60% to an impressive 90.8696%. We also avoided over 162 kg of CO₂ emissions within a single month—an outcome with substantial environmental implications. These results validate the economic and ecological value of integrating predictive intelligence with operational flexibility.

More importantly, our methodology is scalable, replicable, and adaptable to real-world deployments. With modest investments in forecasting tools and battery infrastructure, Repsol can transform energy underutilization into strategic advantage, support national decarbonization goals, and strengthen its leadership in clean energy innovation.

Ultimately, this project moves beyond simulation—it lays the groundwork for real, measurable impact. Our approach proves that with the right data, models, and constraints, industrial sustainability is not only achievable, but also strategically advantageous.