Energy Prediction for Cooperatives: Peak Detection via Classification in Portugal and Output Forecasting via Regression in Ireland

Anny Faria, x24124770

Abstract

This study analyses energy production forecasting and peak demand classification using real-world data from energy communities in Ireland (2020) and Portugal (2022). The datasets include energy meter readings (Wh) and local weather conditions. Two regression models Decision Tree and K-Nearest Neighbors (KNN) are used to predict energy production in Ireland, while XGBoost and an ensemble of Random Forest with Logistic Regression classify peak events in Portugal. The KDD methodology guides the process, with RandomizedSearchCV and cross-validation conducted 10 times used for tuning and validation. To address class imbalance, SMOTEENN is applied. Regression model are evaluated using R^2, MAE and RMSE while Classification models are evaluated using standard metrics like confusion matrices, accuracy, precision, recall, and F1-score. For Regression analysis, KNN achieved better results while in classification analysis XGBoost achieved better results.

Index Terms

Energy Consumption, Energy Production, Regression Models, Classification Models.

I. Introduction

IT is very common researchers predict production from variables such as irradiance, temperature, cloudiness. Research project makes use of a solid KDD workflow, to manage and analyse a rich data set of energy communities in Ireland (2020) and Portugal (2022). This set includes hourly readings of energy meters (Wh) and local meteorological variables, which feed three fundamental analytical fronts: production forecast, classification of peak demand events and two experiments that compared between regression and classification models.

In the first stage, two regression models Decision Tree regression and K-Nearest Neighbors regression (KNN) are applied to predict energy production in Ireland. Then, in the study on Portugal, XGBoost and an ensemble of Random Forest with Logistic Regression are used to classify peak demand events. The entire pipeline follows the KDD methodology, with hyperparameter adjustment via RandomizedSearchCV and ten times repeated cross-validation, ensuring robustness in the selection of models. To reduce class imbalance, SMOTEENN is used, while the results are evaluated by R^2, MAE and RMSE (in regression) and by matrix of confusion, accuracy, precision, recall and F1-score (in classification).

By applying each stage of data selection, pre-processing, exploratory data analysis (EDA), implementation and evaluation model, a reliable flow of data is ensured to generate valuable insights for the optimization of energy management strategies

A. Research Question 1

"How effectively can we predict the peak energy consumption on a short temporal scale applying the hybrid approach SMOTEENN on two machine learning models: XGBoost and a combined Logistic Regression with Random Forest using real-world data collections, and how does hyperparameter optimization with 10-fold cross-validation impact precision, recall, and F1-score?"

B. Research Question 2

"How effectively can we predict the energy production on a shirt temporal scale applying two machine learning models: Regression Tree and K-Nearest Neighbors (KNN) using real-world data collections, and how does hyperparameter optimization with 10-fold cross-validation impact the metrics R^2, MAE, RMSE?

C. Research Outlines

The article is organized as follows.

In Sec. I show a brief overview and motivation about the project point the principal research question. In Sec. II show a wide scope of important knowledge and understanding of the methods used across different domains that serves as a substantial basis for the construction of this project and the research gap. In Sec. III is presented a brief explanation of the choice of methods used along this research project. In Sec. IV show all processes used in detail. In Sec. V present the results comparison beteween models. Finally, in Sec. VI show the reflection and conclusions about the entire project.

II. Literature Review

This section contains a review of relevant work related to machine learning in production and energy consumption to justify the validity of this study.

The prediction of production and energy consumption has been widely studied in the literature and has become the main focus of researchers in electrical and electronic engineering.

Fu and Zhou et al. investigate the peak time for electricity is on the day and hour. using ensemble models. This study adopts the ensemble Random Forest and Gradient Boosting are two tree-based machine learning methods. They choose this model to like categorical vs. continuous. Python is the programming language used to develop the models [1].

Priyadashini and Sahu et al. analyse energy usage in smart homes and demonstrate that predicting power consumption is achievable through machine learning models like Decision Trees (DT), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and k-Nearest Neighbors (KNN). They also propose an ensemble model that combines DT, RF, and XGBoost to assess energy consumption and and compare it to each individual algorithm. Experimental results show that DT-RF-XGBoost ensemble achieved an R^2 approximately 0.99 [2].

Pokharel et al. presented a comparative energy use in a low-energy home in Belgium by examining both indoor and outdoor environmental factors. Forecasting the properties of total energy demand is achieved by training Extreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM). Among them, XGBoost delivers the strongest results on the test set, achieving an R^2 of 0.61, an RMSE of 65.28, an MAE of 29.81, and a MAPE of 28.55 [3].

Banga et al. investigate the detection of electricity theft in smart grids using classification methods. To handling withimbalanced consumption data, six resampling strategies are applied: SMOTE, ADASYN, Random Oversampling, SVM-SMOTE, SMOTEENN (Edited Nearest Neighbor), and SMOTETomek Links. The evaluation proceeds in two phases: first benchmarking twelve individual classifiers, then testing two ensemble approaches. The findings reveal that pairing SMOTEENN with a stacking ensemble produces the highest detection performance of all methods examined [4].

A. Research Gap

No prior work directly contrasts energy consumption and production in local and cooperative energy in two different countries.

III. Choice of Methods

Data preparation is systematically based on the KDD method due to its ability to transform raw data into actionable information. To address highly class imbalance issues in the dataset, the SMOTEENN technique is implemented, especially in datasets about energy that have a lot of null values. It is a reasonable technique to make the dataset more balanced and reduce noise, potentially improving model performance. For the machine learning model, different techniques were chosen to address the specific challenges of energy consumption and production forecasting. In Experiment 1, focused on energy production prediction, regression models such as Decision Tree and K-Nearest Neighbors (KNN) were used due to their ability to model non-linear relationships influenced by climatic variables. In Experiment 2, aimed at classifying peak energy consumption events, ensemble-based models such as Random Forest combined with Logistic Regression and XGBoost were chosen for their robustness in handling class imbalance and detecting critical consumption peaks. While energy consumption studies are increasingly common, predictive modeling of energy production in community-based energy systems remains a relatively unexplored area, reinforcing the importance of this study.

IV. Methodology

This study is based on the concept of discovering knowledge in databases. (KDD) methodology, guiding the process through five main stages. Below are the details of these steps.

A. Data Selection

This study uses two real-world datasets in CSV files based on scientific research articles for two experiment purposes. The first dataset is used to predict total energy production based on weather and weather variables (using two different regression models). The second is used to predict the peak energy consumption based on weather and time variables (using two different classification models).

Dataset 1: *Comprehensive Dataset on Electrical Load Profiles for Energy Community in Ireland*

The focus of this dataset is on local weather parameters and household energy (Wh) from 20 different residences, but for this research purpose, the dataset was filtered from 10 residences' electricity. In this dataset, there are several fields that play a crucial role in capturing, categorizing, and quantifying data related to energy. These include active power consumption, PV generation, grid import and export, charging and discharging, and the state of charge of energy storage. In addition, it gives weather data for the location at a temporal resolution of 1 minute for 2020. The data was acquired directly from the StoreNet project [1].

Dataset 2: *Electricity consumption dataset of a local energy cooperative*

This dataset contains information of the energy consumption measurements of 172 different buildings which are geographically close to each other at Loureiro, Portugal and that that communicate to smart meters every 15 min the amount of energy consumed. The consumption values of one building are related to each column except for the 'Time' and meteorological data. In this context, the original data covers the period between 05–05–2022 and 02–09–2023, but for this research purpose, the dataset was filtered from 15–05–2022 to 15–07–2023. The data includes key attributes such as time, total energy consumption, average air temperature, total global radiation, and various climate attributes. Some readings for certain buildings are missing and are marked as NaN. To enrich the energy measurements, we incorporated local weather data recorded at the same times. Both the energy and meteorological datasets share identical timestamps and row counts. The weather observations were obtained from the station closest to Loureiro. The data was acquired directly from data mendeley [2].

B. Data Pre-processing

Several pre-processing steps were applied to ensure data quality and readiness for modeling:

The purpose of the initial exploratory data analysis is to gain knowledge about the datasets structure and variables, basic statistics, completeness, and unique values in categorical or object features. The data cleaning and feature engineering processes are guided by valuable insights provided by this step.

For Dataset 1: To addressing null, missing, and duplicated values involves either imputing or removing them as part of the cleaning process. For example, the dataset 1 used in this research has no duplicated values. However, the feature, `rain`, contains empty strings (`' '`) and 0 values.

The dataset is prepared for machine learning models by implementing the data feature engineering and transformation process. In the dataset 1, the following transformations were made: date column such as `date`, were converted to datetime type. Some new features were created: from the column `date` were created features that reveal specific seasonalities, trends, or behaviors like `Date`, `Month`, `Hour`, `Minute`, `Day_of_Week` and `Weekend`. None of the columns were encoded using encoding because the dtype of the dataset 1 contains floats64 values.

Outliers in numerical columns `Total Consumption(Wh)`, `speed`, `drybulbsbl` and were treated using the interquartile range (IQR) method [5]. The integrity of the system is maintained while extreme outliers are effectively managed by this approach. dataset 1. The presence of outliers increases the variability within the datasets, potentially reducing the statistical power. By extracting and removing outliers, the significance and validity of the results can be improved. I The study focus was on eliminating outliers to avoid duplication and produce standardized data for the processing and data mining phase.

For Dataset 2: The cleaning process involves imputing missing values with the average. In this context, columns with 172 `Energy_Meter_` are filled by picking 5 "neighbor" meters, computing their average, and filling NaNs in this meter with that average [6].

The data feature engineering and transformation process were made: date column such as `Time`, were converted to datetime type. Some new features were created: from the column `Time` were also created those features that reveals specific seasonalities, trends, or behaviors. From the column `Total_Energy_Consumption` were derived `Lag_1H` and `Lag_2H` corresponding to one-hour and two-hour shifts. Other features related to temperature and climate conditions were created. The target `Peak_Consumption` was created by calculating the 75th percentile of the consumption for each hour. The column `Season` is encoded using "dummies" because the dtype is object.

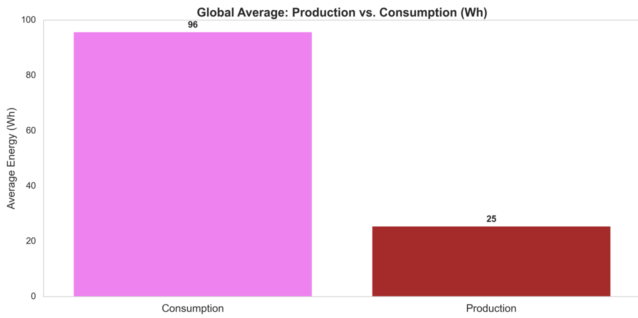
Outliers in numerical columns `Avg_Temp`, `Avg_Rel_Humidity`, `Avg_Wind_Speed`, `Max_Inst_Wind_Speed`, `Inst_Temp`, `Total_Global_Rad` and `Total_Energy_Consumption` were treated using the interquartile range (IQR) method.

C. Exploratory Data Analysis (EDA)

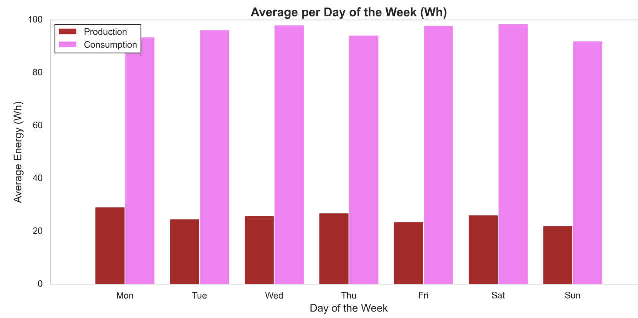
In the EDA process, different aspects of the energy consumption and production are explored based on the weather and time variables (date, hour, weekdays, etc).

For Dataset 1:

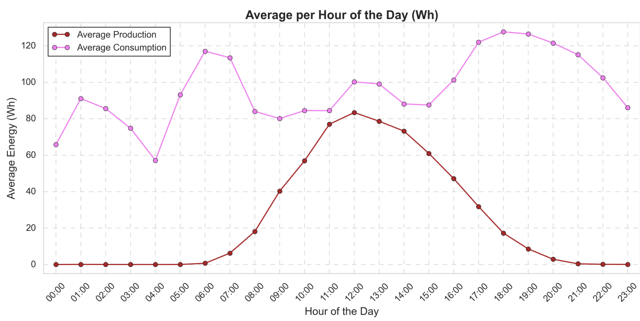
[1]https://www.nature.com/articles/s41597-024-03454-2/tables/7
[2]https://data.mendeley.com/datasets/vryvyfz2tj/1



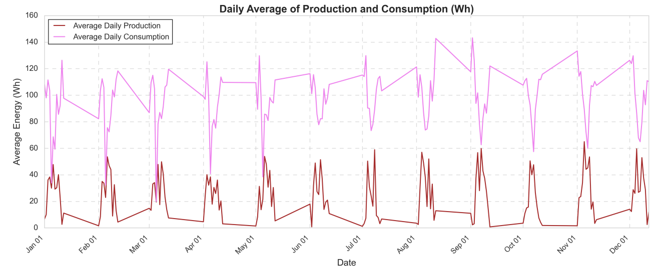
(a)



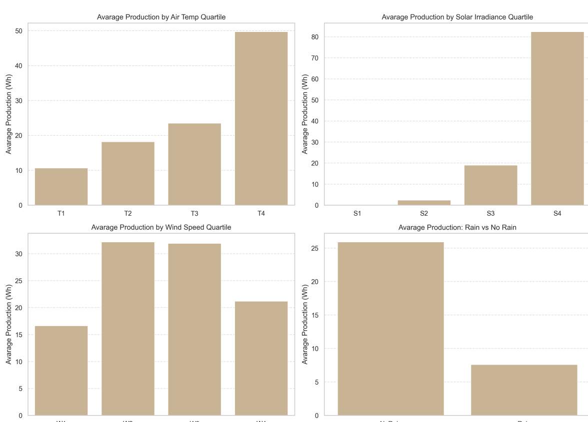
(b)



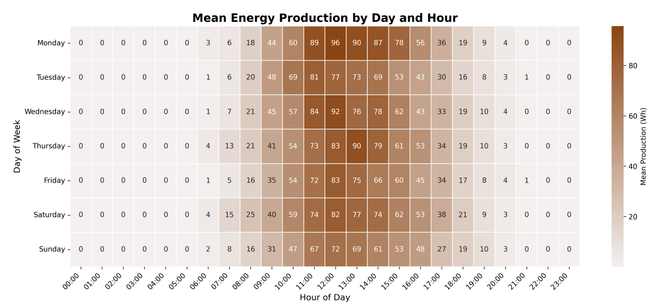
(c)



(d)



(e)



(f)

Fig. 1: Dataset 1: Exploratory data analysis visualizations

The fig. 1a demonstrates that the average energy deficit production only covers about 26% of consumption.

The fig. 1b shows that the consumption remains relatively constant between days. But production varies during the weekdays with Monday being the best day of energy production.

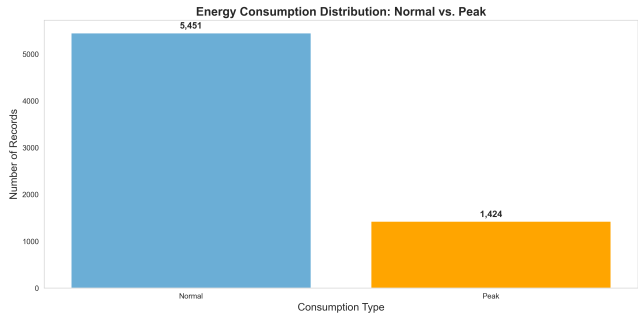
The fig. 1c reveals that the average energy consumption is high throughout the day, with peaks at 06:00h, 12:00h and between 18:00h-22:00h. On the other hand, the average energy production only start 7:00h and decreases at 13:00h, between 09:00h-16:00h the energy production partially covers energy consumption.

The fig. 1d shows that energy production has a seasonality behavior that influences energy consumption and production along the year.

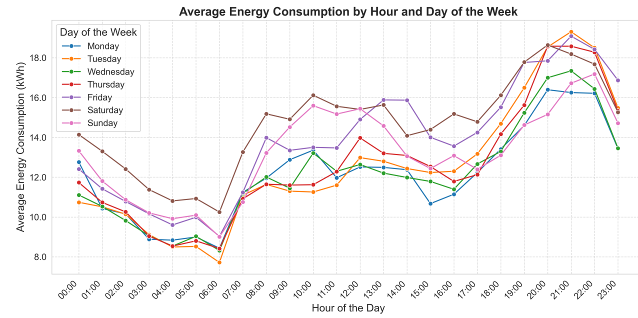
The fig. 1e shows four different graphs (i)-(iv) of the average energy production in relation to climatic events (temperature, solar radiation, wind speed, and rain). The first graph (i) indicates that temperature increases energy production and reaches a peak in T4 of approximately 50 (Wh). The second (ii) implies that energy production is highly effective in conditions of high solar radiation S4, with an energy production slightly higher than 80 (Wh). In the third (iii), it shows that energy production is high in W2-W3 and decreases in W4, suggesting a potential negative impact on energy production in very strong wind conditions. The fourth (iv), It shows that the average production on days without rain is about 26 (Wh), while on days with rain it can vary by more or less 8 (Wh). In general, the climatic factor that is most important for energy production is solar radiation. Rain and very strong wind reduce production.

The fig. 1f Production occurs between 6:00h and 20:00h, with peak between 11:00h and 14:00h. Also, monday presents the highest peak 96 (Wh) at 12:00h. Production is zero during the night 00:00h-05:00h and after 21:00h, showing that energy production has strongly relation with sunlight.

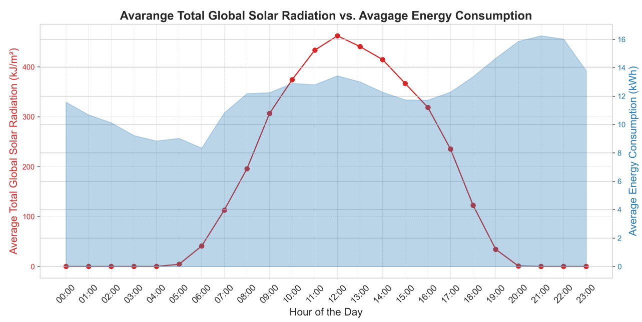
For Dataset 2:



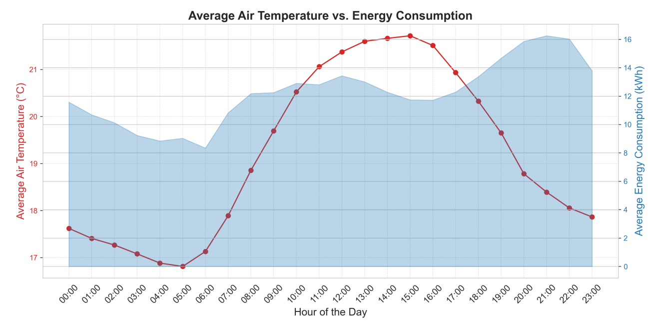
(a)



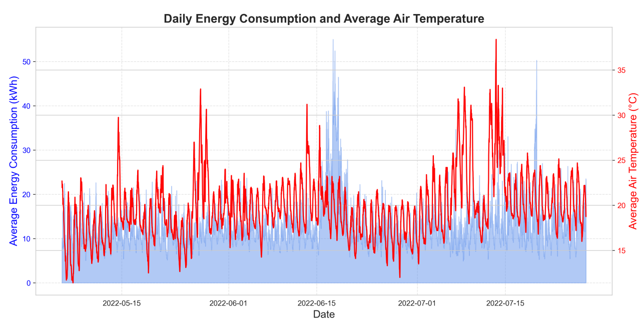
(b)



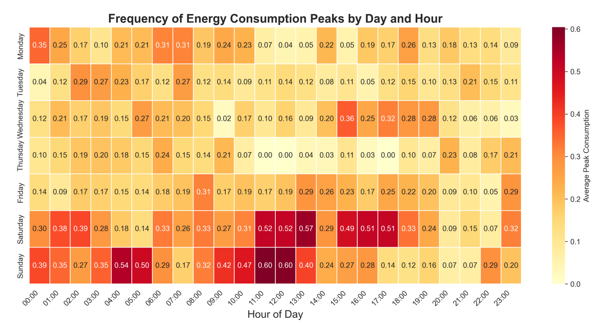
(c)



(d)



(e)



(f)

Fig. 2: Dataset 2: Exploratory data analysis visualizations

The fig. 2a shows that most records are at a normal consumption level of approximately (5451 counts), but there is a relevant volume (1424 count) of energy peaks.

The fig. 2b reveals night peaks around 20:00h-22:00h for every day. On weekends, Saturday and Sunday have higher consumption in the daytime compared to the week. The lowest energy consumption is between 02:00h and 06:00h.

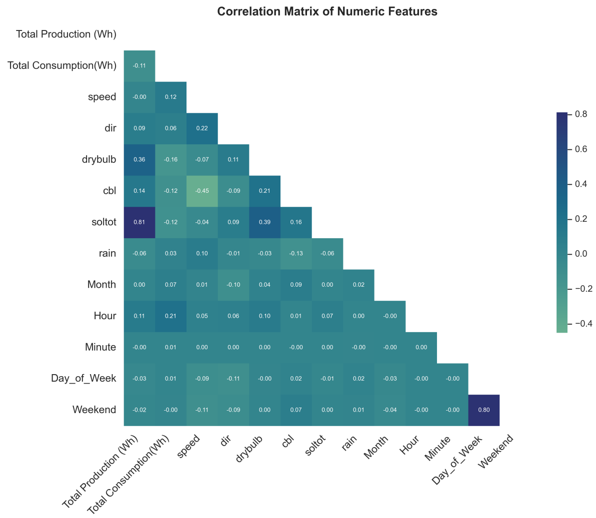
The fig. 2c suggest that the average consumption per hour with peak radiation is between 11h and 14h. Already consumption grows at the end of the day. Energy consumption does not grow linearly with solar radiation.

The fig. 2d shows temperature rises from 06:00h to 14:00h and falls after that. Consumption, on the other hand, has two peaks: at 09:00 and another between 19:00-22:00.

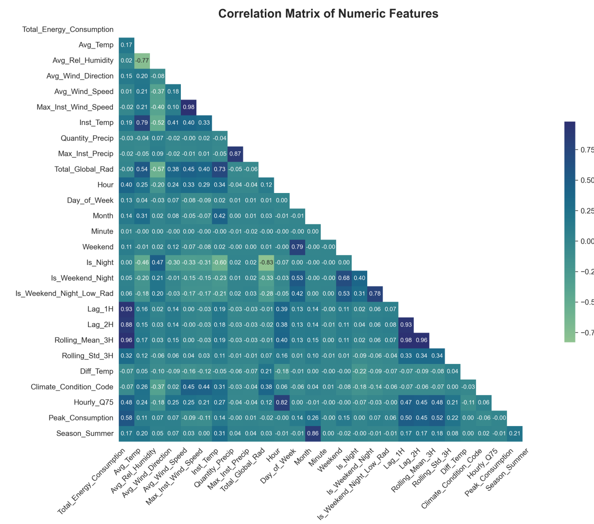
The fig. 2e shows an analysis of energy consumption between May and July. The temperature in this period rises over time until the peak of summer. The consumption scales enough with frequent peaks but does not grow linearly with temperature.

The fig. 2f reveals the most critical times for energy consumption, so weekends are more prone to spikes.

Lastly, a correlation analysis was performed to explore relationships between numerical variables. This analysis highlighted potential correlations. By identifying potential redundancies in the data and offering valuable insights into customer behavior, these findings offer valuable insights.



(a) Dataset 1: Correlation Heatmap Analysis.



(b) Dataset 2: Correlation Heatmap Analysis.

Fig. 3: Visual representation of the correlation between variables for each dataset.

The fig. 3a suggests a very strong correlation (0.81) between the variables production and solar radiation (`soltot`). Also, we have a moderate correlation (0.36) between production and temperature (`drybulb`). There is a weak but positive correlation (0.12) between the variable production and time which is consistent with the daily solar production cycle. Finally, the variable rain (`rain`) has a negative correlation with production (-0.06). Overall, solar radiation is the best predictor of energy production.

The fig. 3b suggests that a more strongly correlated variable with peak consumption is a recent consumption variability (`Rolling_Std_3H`), followed by the mobile media (`Rolling_Mean_3H`). The time variable (`Hour`) also has a strong correlation (0.26) with total energy consumption. Climatic factors have a moderate influence on consumption.

D. *Experimental Setup & Model Implementation*

After the completion of all the pre-processing steps, the datasets for experiment 1 and experiment 2 were split into training (80%) and testing (20%) sets while isolating the target variable.

Experiment 1: In this experiment, we analyse the performance of two distinct machine learning models: Regression Tree and KNN Regression. In addition, the feature standardization was applied using StandardScaler to ensure uniform input scaling before training and testing the model.

For the Regression Tree, the best hyperparameters were determined to optimize the model using RandomizedSearchCV with the number of interactions (n_int) set at 40, time series cross-validation is used with n_splits set at 5, and the related metric evaluation (score) negative mean squared error is as follows: the maximum depth of trees (max_depth) was set to 12, and the number of features considered when looking for the best split (max_features) was set to 'None.' The minimum number of samples required to split an internal node (min_samples_split) was set to 9, the minimum number of samples required to split an internal node (min_samples_split) was set to 6, the models are trained and tested on the same data (random_state) was set at 42. In addition, the scoring metric used in RandomizedSearchCV was negative mean squared error [3], but it was later converted to RMSE for evaluation to better interpretation.

For KNN Regression, the hyperparameter tuning was set the nearest training samples used to make predictions (n_neighbors) for a range of randint(1, 30). The time series cross-validation is used with n_splits set at 3. The number of interactions (n_int) set at 30 different values of k. Also the related metric avaliation (score) was based on negative mean squared error. The models are trained and tested on the same data (random_state) was set at 42. The best k found was 29.

Experiment 2: In this experiment we analyze the performance of two distinct machine learning models: XGBoost and combined ensemble (Random Forest + Logistic Regression). Initially, the dataset exhibited significant class imbalance between class 0 (Normal) and class 1 (Peak). To address this issue, the SMOTEENN technique was applied stratified over the training set, resulting in a balanced distribution for class 0 and class 1, as shown in Tab. I.

Peak Consumption	Before SMOTEENN	After SMOTEENN
0 (Normal)	6000	2589
1 (Peak)	2000	1822

TABLE I: Class distribution before and after SMOTEENN

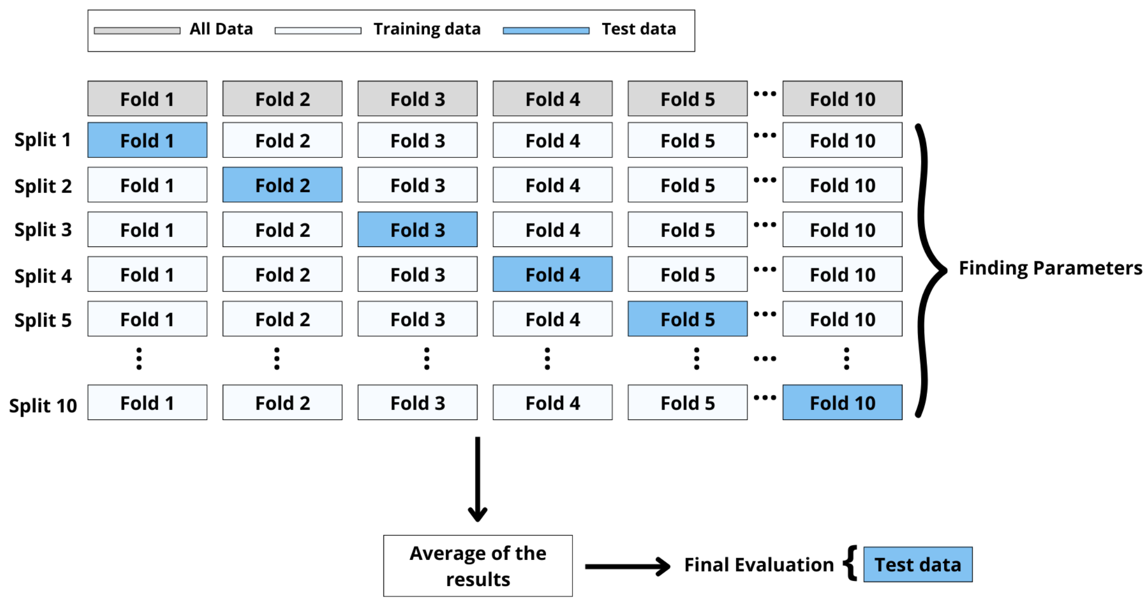
[3]https://stackoverflow.com/questions/48244219/is-sklearn-metrics-mean-squared-error-the-larger-the-better-negated

For XGBoost model, was made an XGBoost model to be trained with balanced data balancead to capture relation beteween variables. In this context scale_pos_weight was set 3 to deal with imbalanced class, the number of trees (n_estimators) was set at 100, the evaluation metrics for validation data (eval_metric) was set as logloss, and the models are trained and tested on the same data (random_state) was set at 42.

Subsequently, feature importances are provided by the fitted attribute (feature_importances_) and only features with importance > 0 were selected. Then, balanced data was set to keep only top features. This technique was made to guarantee that the next models be trained only with top features.

The XGBoost model was set again and the best hyperparameters were determined using RandomizedSearchCV with number of interactions (n_int) set at 50, 10-fold cross validation (cv), metric avaliation (score) f1 as follows: the number of trees (n_estimators) was set at 300, the maximum depth of trees (max_depth) was set to 5, and the number of features considered when looking for the best split (colsample_bytree) was set to 0.8, the learning rate was set to 0.1, the gamma parameter was set to 0.3, the subsample ratio of the training instances was set to 0.6, the L2 regularization term (reg_lambda) was set to 1, the L1 regularization term (reg_alpha) was set to 0, and the models are trained and tested on the same data (random_state) was set at 42.

The fig. 4a shows the performance of $10-$fold cross validation based on scikit$-$learn guide [4]



(a) Diagram to better illustrate the model performance in $10-$fold cross-validation.

For the ensemble model combining Random Forest + Logistic Regression, was constructed using a soft voting strategy. The base classifiers included a RandomForestClassifier (random_state = 42) and a LogisticRegression model configured with a maximum number of iterations (max_iter) set to 5000. The best hyperparameters were determined using RandomizedSearchCV with number of interactions (n_int) set at 50, 10-fold cross validation (cv), metric avaliation (score) f1 as follows: the number of trees in the Random Forest (rf_n_estimators) was set at 257, the maximum depth of trees (rf_max_depth) was set to None, and the regularization strength for the Logistic Regression (lr_C) was set to 0.5806. The models are trained and tested on the same data (random_state) was set at 42.

After tuning for experiment 1 and experiment 2, the best performing ensemble model was selected and used for final evaluation.

[4]https://scikit- learn.org/stable/modules/cross_validation.html

V. Model Evaluation

For experiment 1, the standard classification metrics were chosen: R^2 (Coefficient of determination), MAE (Mean absolute error), RMSE (Root Mean Squared Error) , defined as follows:

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad \mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

where $\hat{y}$ is the predicted value of y and $\bar{y}$ is mean value of y, n is the size of sample.

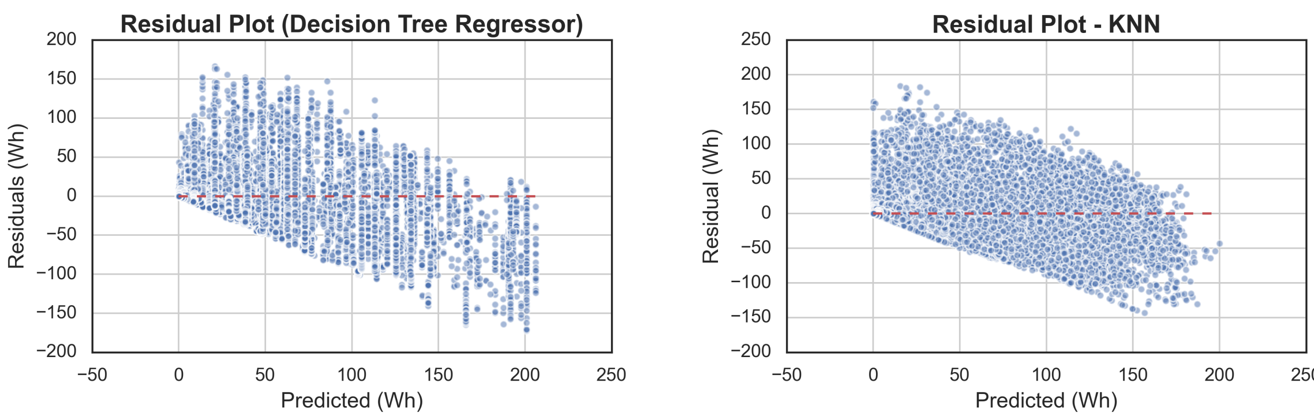
For experiment 2, the standard classification metrics were chosen: accuracy, precision, recall, and F1-score, defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}, \quad \text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}, \quad \text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative. respectively.

For experiment 1 and experiment 2, as mentioned in the previous section, the F1 score was chosen to evaluate the metrics because, as we can see on the F1 score equation, it can be interpreted as a weighted average of precision and recall, where the relative contributions of precision and recall to the F1-score are equal. What we are trying to achieve with the F1-score metric is to find an equal balance between precision and recall, which is extremely useful for the imbalanced datasets that we are working with.

A. *Experiment 1: Comparison Models*



(a) Residual plot of Decision Tree Regressor.

(b) Residual Plot of KNN Regressor.

Fig. 5: Visual representation of the residual plot for each model.

Metric	Regression Tree	KNN	Best Performance
R^2	0.58	0.63	KNN
MAE	13.30	14.00	Regression Tree
RMSE	28.45	26.74	KNN

TABLE II: Performance comparison between Regression Tree and KNN models

Table II presents the performance comparison between two regression models (Regression Tree (Regression Tree) and KNN Regressor), based on three metrics: R^2, MAE and RMSE. The KNN model obtained a value of $R2 = 0.63$, higher than the value of 0.58 of the regression tree. This indicates that the KNN is more effective in explaining the variability of energy production by better capturing the global patterns of the data set. On the other hand, the regression tree presented the lowest MAE (13.30), better than the MAE of the KNN (14.00). This suggests that, on average, the tree's forecasts are a little closer to the actual values, although the difference is not so great. KNN obtained an RMSE of 26.74, lower than the regression tree (28.45), which demonstrates that the KNN model makes fewer large errors.

In summary, the KNN model presents better overall performance, with greater generalization capacity and lower sensitivity to large errors. The regression tree, despite its good MDD, presents greater variability in errors.

B. Experiment 2: Comparison Models

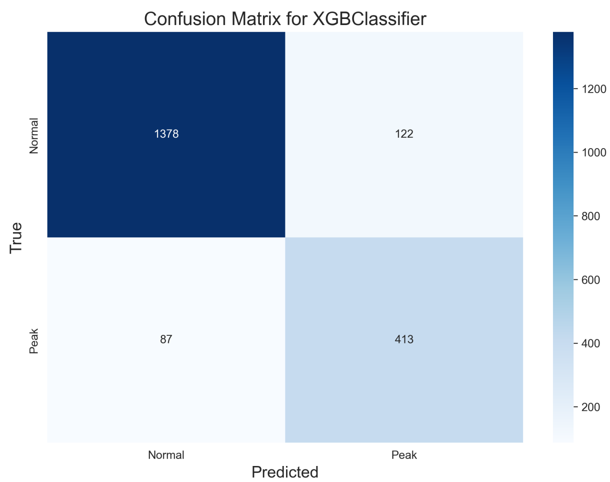
Metric	XGBoost	Ensemble (RF + LR)	Best Performance
Accuracy	0.90	0.89	XGBoost
Precision (Class 1 – Peak)	0.77	0.77	Both
Recall (Class 1 – Peak)	0.83	0.79	XGBoost
F1-Score (Class 1 – Peak)	0.80	0.78	XGBoost

TABLE III: Performance comparison between XGBoost and Ensemble (Random Forest + Logistic Regression)

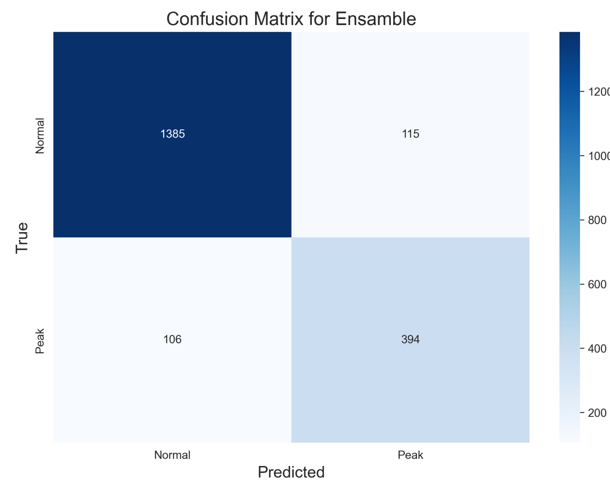
Table III shows the performance comparison between two classification approaches XGBoost and an ensemble combining Random Forest with Logistic Regression on the task of detecting peak events (class 1). The metrics evaluated are Accuracy, Precision, Recall and F1-Score.

XGBoost achieved a higher accuracy (0.90) than the ensemble (0.89), indicating that it makes a greater proportion of correct predictions overall. Both models attained the same precision of 0.77, meaning they correctly flagged 77% of their predicted peaks. However, XGBoost also outperformed the ensemble in recall (0.83 vs. 0.79), demonstrating greater sensitivity in capturing actual peak events. As a result of this The ensembles F1-Score was 0.78, but XGBoost achieved an F1-Score of 0.80 because of the better balance between precision and recall.

In summary, across all four metrics XGBoost outperforms the combined Random Forest + Logistic Regression model and is therefore the preferred choice for peak detection in energy production.



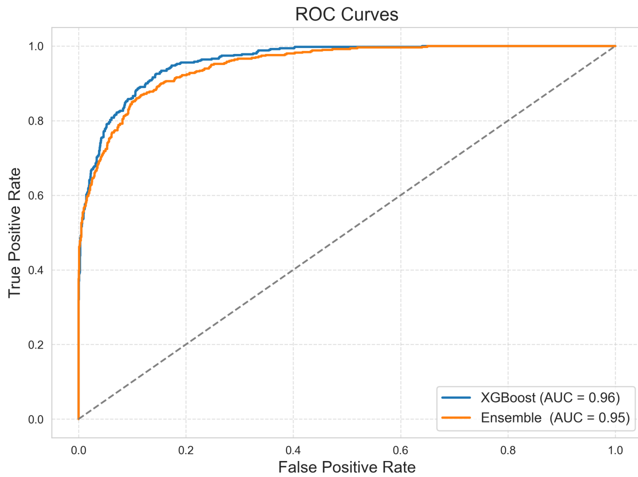
(a) Confusion matrix for XGBoost.



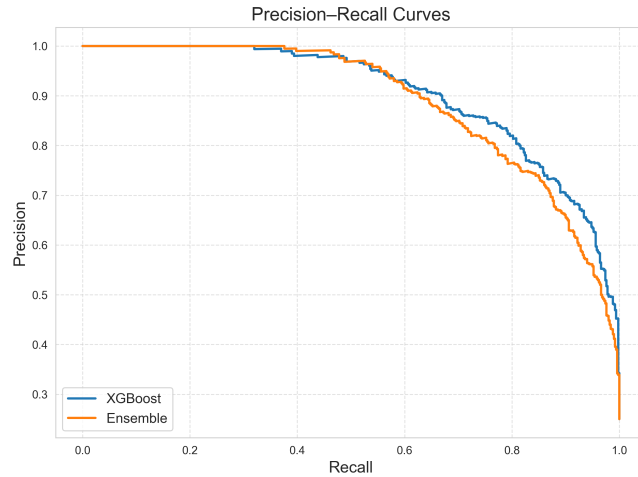
(b) Confusion Matrix for Ensamble.

Fig. 6: Visual representation of the confusion matrix for each model.

The Fig. 6 shows the differences in their classification behavior. The Ensemble model correctly classified 1385 normal instances and misclassified 115 normal instances as peaks, while XGBoost correctly classified 1378 normal cases and misclassified 122. For peak detection, the Ensemble model correctly identified 394 peak instances but failed to detect 106, whereas XGBoost correctly identified 413 and missed only 87. This means XGBoost was more effective at detecting true peaks with a higher true positive count and a lower false negative count. On the other hand, the Ensemble model produced less false positives.



(a) ROC AUC curve comparison.



(b) Precision curve comparison.

Fig. 7: Curve comparison.

The Fig. 7 shows the precision–recall and ROC curves that provides the classification performance of the XGBoost and Ensemble models.

In the ROC curve, both models demonstrate strong performance with areas under the curve (AUC) close to 1. XGBoost achieves an AUC of 0.96, while the ensemble model reaches 0.95. This reinforces that both classifiers are effective at distinguishing between peak and normal classes. However, the XGBoost curve stays slightly above the Ensemble across most of the plot.

The precision–recall curve shows that the XGBoost model has higher precision than the Ensemble model. This means that for a given recall, XGBoost tends to produce fewer false positives.

In general, XGBoost slightly outperforms the Ensemble model in both precision–recall space and ROC space.

VI. Conclusions and Future Work

In this project, different machine learning models were selected to address the specific challenges of energy consumption and production forecasting. In Experiment 1, focused on the prediction of energy production, regression models such as Decision Tree and K-Nearest Neighbors (KNN) were used due to their ability to model non-linear relationships influenced by climatic variables. In Experiment 2, aimed at the classification of peak events in energy consumption, ensemble-based models were chosen, such as Random Forest combined with Logistic Regression and XGBoost, for its robustness in the treatment of unbalanced classes and in the detection of critical consumption peaks. To ensure a reliable and impartial evaluation of the models, the 10-fold cross-validation technique was applied in all experiments. Although studies on energy consumption are becoming more common, the predictive modeling of energy production in local energy communities is still a relatively unexplored area, which reinforces the relevance of this work.

For future projects, it would be possible to apply the solar radiation values from these datasets to a photonic material in order to predict the wavelength range in which light will propagate through the material. Through this, we may be able to forecast which materials absorb solar energy more efficiently, ultimately contributing to improved energy production in residential settings.

References

[1] T. Fu et al., "Predicting peak day and peak hour of electricity demand with ensemble machine learning," Frontiers in Energy Research, vol. 10, p. 944804, 2022.

[2] Priyadarshini, S. Sahu, R. Kumar and D. Taniar, "A machine-learning ensemble model for predicting energy consumption in smart homes," Internet of Things, vol. 20, p. 100636, 2022.

[3] S. Pokharel and P. Ghimire, "Data-driven ML models for accurate prediction of energy consumption in a low-energy house: A comparative study of XGBoost, Random Forest, Decision Tree, and Support Vector Machine," Journal of Innovations in Engineering Education, vol. 6, no. 1, pp. 12–20, 2023.

[4] A. Banga, R. Ahuja and S. C. Sharma, "Accurate detection of electricity theft using classification algorithms and Internet of Things in smart grid," Arabian Journal for Science and Engineering, vol. 47, pp. 9583–9599, 2022.

[5] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," in Information and Decision Sciences: Proceedings of the 6th International Conference on FICTA, Singapore: Springer, 2018, pp. 511–518.

[6] T. Decorte, S. Mortier, J. J. Lembrechts, F. J. Meysman, S. Latré, E. Mannens and T. Verdonck, "Missing value imputation of wireless sensor data for environmental monitoring," Sensors, vol. 24,, p. 2416, 2024.

[7] R. Trivedi, M. Bahloul, A. Saif, S. Patra and S. Khadem, "Comprehensive dataset on electrical load profiles for energy community in Ireland," Scientific Data, vol. 11, no. 1, p. 621, 2024.

[8] F. Monteiro, R. Oliveira, J. Almeida, P. Gonçalves, P. Bartolomeu, J. Neto and R. Deus, "Electricity consumption dataset of a local energy cooperative," Data in Brief, vol. 54, p. 110373, 2024.

[9] G. K. Tso and K. K. Yau, “Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks,” Energy, vol. 32, no. 9, pp. 1761–1768, 2007.

[10] P. Nie, M. Roccotelli, M. P. Fanti, Z. Ming and Z. Li, “Prediction of home energy consumption based on gradient boosting regression tree,” Energy Reports, vol. 7, pp. 1246–1255, 2021.