future internet

MDPI

Article

Combining Advanced Feature-Selection Methods to Uncover Atypical Energy-Consumption Patterns

Lucas Henriques [1,2,*], Felipe Prata Lima [3] and Cecilia Castro [1,*]

[1] Centre of Mathematics, Universidade do Minho, 4710-057 Braga, Portugal
[2] Mathematics Department, Federal Institute of Alagoas, Maceió 57020-600, AL, Brazil
[3] IT Department, Federal Institute of Alagoas, Maceió 57020-600, AL, Brazil; felipepratalima@gmail.com
* Correspondence: lucasdestefano2@hotmail.com (L.H.); cecilia@math.uminho.pt (C.C.)

Abstract: Understanding household energy-consumption patterns is essential for developing effective energy-conservation strategies. This study aims to identify 'out-profiled' consumers—households that exhibit atypical energy-usage behaviors—by applying four distinct feature-selection methodologies. Specifically, we utilized the chi-square independence test to assess feature independence, recursive feature elimination with multinomial logistic regression (RFE-MLR) to identify optimal feature subsets, random forest (RF) to determine feature importance, and a combined fuzzy rough feature selection with fuzzy rough nearest neighbors (FRFS-FRNN) for handling uncertainty and imprecision in data. These methods were applied to a dataset based on a survey of 383 households in Brazil, capturing various factors such as household size, income levels, geographical location, and appliance usage. Our analysis revealed that key features such as the number of people in the household, heating and air conditioning usage, and income levels significantly influence energy consumption. The novelty of our work lies in the comprehensive application of these advanced feature-selection techniques to identify atypical consumption patterns in a specific regional context. The results showed that households without heating and air conditioning equipment in medium- or high-consumption profiles, and those with lower- or medium-income levels in medium- or high-consumption profiles, were considered out-profiled. These findings provide actionable insights for energy providers and policymakers, enabling the design of targeted energy-conservation strategies. This study demonstrates the importance of tailored approaches in promoting sustainable energy consumption and highlights notable deviations in energy-use patterns, offering a foundation for future research and policy development.

Keywords: behavior analysis; consumption patterns; feature selection; fuzzy rough sets; random forest

Citation: Henriques, L.; Lima, F.P.; Castro, C. Combining Advanced Feature-Selection Methods to Uncover Atypical Energy-Consumption Patterns. *Future Internet* **2024**, *16*, 229. https://doi.org/10.3390/fi16070229

Academic Editor: Gianluigi Ferrari

Received: 2 June 2024
Revised: 18 June 2024
Accepted: 25 June 2024
Published: 28 June 2024

1. Introduction

Household energy conservation is a critical issue with significant economic and environmental implications. Reducing energy usage not only helps families lower their energy bills but also minimizes the demand for energy generation, leading to broader benefits such as decreased environmental impact and enhanced economic sustainability. Previous studies have emphasized the importance of understanding both subjective and objective factors influencing household energy consumption [1,2].

Identifying 'out-profiled' consumers—households exhibiting atypical energy-usage behaviors—is crucial for optimizing energy distribution, planning for energy demand and implementing effective energy-saving measures [3,4]. These atypical behaviors can include significantly higher or lower energy usage than similar households, inconsistent consumption patterns or unexpected peaks in energy usage. For instance, a household may have a high number of occupants but exhibit lower than expected energy consumption, or vice versa [5]. There is a need for more refined methods to accurately identify and understand these unique consumption patterns, providing actionable insights for energy

providers and policymakers and enabling the development of targeted strategies for energy conservation tailored to specific household dynamics [6].

Various studies have explored different methods for analyzing energy-consumption patterns, which can be broadly categorized into methods for typical and atypical energy-consumption patterns. Several innovative methods have been proposed to analyze energy-consumption patterns. For instance, ref. [7] introduced a deep anomaly detection technique using energy time-series images, which provides a robust framework for identifying unusual consumption patterns in buildings. Another study by ref. [8] proposed a data-driven model that leverages statistical and machine learning methods to analyze energy consumption in residential buildings. In Greece, ref. [9] examined the energy-consumption patterns of residential users, offering insights into the factors influencing energy use in different contexts.

Typical energy-consumption patterns have been extensively studied using various forecasting techniques. Ref. [10] utilized a spectral clustering algorithm combined with temporal fusion transformers to forecast building energy consumption. This approach enhances the interpretability of the models used. Similarly, ref. [11] presented a new method for seasonal energy consumption-forecasting employing temporal convolutional networks, which effectively capture the seasonal variations in energy use.

Identifying atypical energy-consumption patterns is a relatively newer area of research. Studies such as ref. [12] have focused on uncovering unusual consumption behaviors in households. Our study advances this field by combining multiple feature-selection techniques, including the Chi-square test, recursive elimination feature, random forest and fuzzy rough feature selection to identify key determinants of atypical energy usage. This comprehensive approach provides a nuanced understanding of household energy consumption, highlighting deviations that can inform targeted energy-conservation strategies.

Feature selection plays a vital role in refining these methods. It is a technique that reduces the data-processing scale by removing irrelevant and redundant features, thereby reducing dimensionality, improving learning accuracy, reducing learning time and simplifying learning results [13,14]. No single feature-selection method is universally applicable to all datasets [15]. Each dataset possesses unique characteristics that influence the effectiveness of different feature-selection techniques. Consequently, it is essential to evaluate and test various methods to determine the most suitable approach for a given dataset.

There are three primary types of feature-selection techniques: filter methods, wrapper methods and embedded methods. Filter methods are computationally efficient because they are independent of the classifier, making them easy and quick to implement [16]. The most common filter methods use statistical measurements to determine correlation or independence between input features and the target variable [17]. Wrapper methods evaluate the features and produce the result simultaneously with a learning model using a learning algorithm, resulting in higher classification accuracy but with the disadvantage of high time complexity [18]. Embedded methods incorporate the classifier's bias into feature selection, producing better classifier performance and greater efficiency since they do not need to evaluate feature sets iteratively [19,20].

This study aims to address the gap in understanding atypical energy-usage patterns by employing advanced feature-selection techniques to uncover the critical determinants of household energy usage. In our research, we tested several feature-selection methods: the chi-square test to assess the features' independence [17], recursive feature elimination (RFE) with multinomial logistic regression (MLR) [18], random forests (RF) [15] and fuzzy rough sets (FRS) [13,14]. These techniques help identify the most informative features affecting energy consumption and offer a deeper understanding of consumer behaviors.

The dataset used in this study is part of a survey administered to 383 randomly selected households in Brazil, collecting data on monthly electrical power consumption and the characteristics of the households and occupants [21]. This comprehensive dataset provides a robust foundation for analyzing energy-consumption patterns and identifying key determinants. The household energy usage was categorized into three distinct classes: low-, medium- and high-load-consumption profiles.

By applying these advanced feature-selection techniques, this research aims to set a new standard in energy conservation, offering a robust template for future studies focused on optimizing residential energy use. The ultimate goal is to empower urban actors to make data-driven decisions for a more sustainable future. Beyond the Introduction, the structure of this paper is as follows: Section 2 outlines the methodologies employed in our study to identify critical determinants of household energy usage. Section 3 provides an overview of the dataset used in this study, including the data-collection process, the characteristics of the dataset and the categorization of variables. Section 4 discusses the feature-selection process, applying specific criteria to obtain a subset of relevant features. Section 5 analyzes the selected features, assessing their impact on household energy load profiles and identifying out-profiled households. Finally, Section 6 concludes the study and suggests future research directions.

2. Methodology

This section outlines the methodologies employed in our study to identify critical determinants of household energy usage. Various feature-selection techniques were utilized to refine our dataset, ensuring that only the most relevant features were retained for analysis. The overall feature-selection process and the interaction between different methods are illustrated in Figure 1.

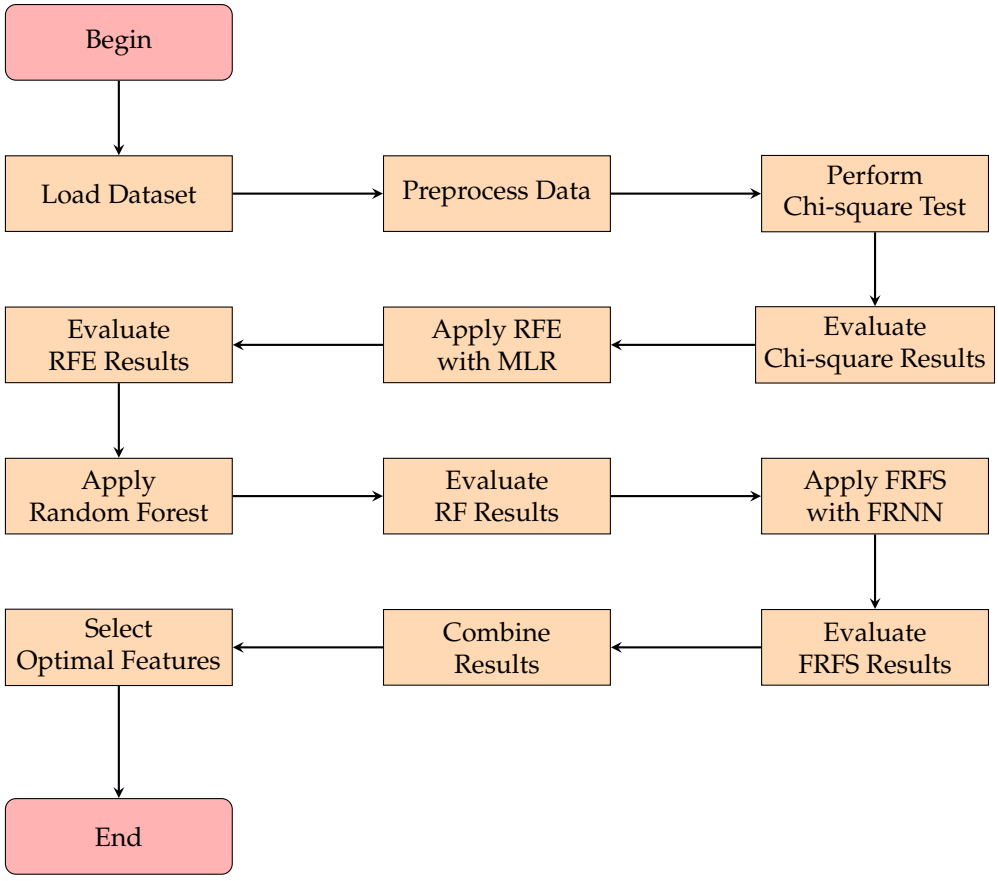


Figure 1. Flowchart of the feature-selection process and interaction between different methods.

2.1. Chi-Square Independence Test

The chi-square (CHI2) test examines whether there is a relationship between two categorical variables in a single sample. This test assesses the independence or association between two categorical variables [22]. The null hypothesis (H_0) suggests that there is no association between the two variables, implying they are independent. Conversely, the alternate hypothesis (H_1) posits that there is a significant association between them,

indicating they are not independent. From a contingency table (Table 1) that displays the frequency counts of the joint occurrences n_{ij} of the two categorical variables, let X represent one variable with categories $X_1, X_2, \ldots, X_m$ and Y represent the other variable with categories $Y_1, Y_2, \ldots, Y_k$.

Table 1. Contingency table for the CHI2 test.

	Variable Y				
	Y_1	Y_2	$\cdots$	Y_k	
X_1	n_{11}	n_{12}	$\cdots$	n_{1k}	x_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2k}	x_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_m	n_{m1}	n_{m2}	$\cdots$	n_{mk}	x_m
	y_1	y_2	$\cdots$	y_k	n

The values x_i and y_j are the marginal totals of the rows and columns, respectively, and are calculated as $x_i = \sum_{j=1}^{k} n_{ij}$, $i = 1, \ldots, m$ and $y_j = \sum_{i=1}^{m} n_{ij}$, $j = 1, \ldots, k$. Under the null hypothesis of independence, the expected frequency (E_{ij}) for each cell in the contingency table is

$$E_{ij} = \frac{x_i y_j}{n}, \tag{1}$$

where $n = \sum_{i=1}^{m} \sum_{j=1}^{k} n_{ij}$.

The chi-square statistic (χ^2) measures the discrepancy between the observed frequencies (n_{ij}) and the expected frequencies (E_{ij}) given by (1), and is defined as

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}.$$

Under certain conditions, such as when the sample size is sufficiently large and the expected frequencies are not too small, the χ^2 statistic follows a chi-square distribution with $(m-1)(k-1)$ degrees of freedom. Typically, each expected frequency E_{ij} should be at least 5 to ensure the validity of the chi-square approximation.

To determine whether to reject the null hypothesis, we use the χ^2 statistic. The null hypothesis is rejected if the χ^2 statistic exceeds the critical value from the chi-square distribution at a chosen significance level (such as $\alpha = 0.05$). Alternatively, the null hypothesis is rejected if the p-value is less than the significance level.

It is important to note that the chi-square test is an omnibus test. Therefore, if the test indicates a significant association, post hoc procedures need to be conducted to compare individual conditions.

2.2. Recursive Feature Elimination with Multinomial Logistic Regression

The RFE process aims to select the most important features by iteratively removing the least significant ones, thereby improving the efficiency of the feature-selection process. As explained in Algorithm 1, the process begins with constructing a model using all available features. Subsequently, each feature is assigned a weight based on its relevance in classifying the target variable. The feature with the lowest weight is then eliminated. The model is subsequently reconstructed, and the importance of each remaining feature is recalculated [23]. This iterative process continues until the remaining set of features meets a predefined performance threshold.

Algorithm 1 Recursive Feature Elimination (RFE)

Input: Dataset X with feature dimension k
Output: $S^* \leftarrow$ Set of features providing the highest performance metric

Initialize: $S \leftarrow \{1, \ldots, k\}$
Train a model on X using features in S
min_score $\leftarrow$ S performance metric
for each feature f in S **do**
  $S' \leftarrow S - \{f\}$
  Train a model on X using features in S'
  if S' performance metric $<$ min_score **then**
    $f_{\text{min}} \leftarrow f$
    min_score $\leftarrow$ S' performance metric
  end if
end for
Remove f_{min} from S

Given that the response variable in this study, household energy usage, is categorized into three distinct classes (low-, medium- and high-load-consumption profiles), we employed a multinomial logistic regression (MLR) algorithm. This generalized linear model is particularly suitable for situations where the response variable encompasses more than two categories. The MLR algorithm utilizes a non-linear log transformation, which facilitates the calculation of the probability of occurrence for each class of the dependent variable [24].

Several assumptions were verified to ensure the applicability of the multinomial logistic regression:

- Independence of irrelevant alternatives: The assumption that the odds of preferring one class over another are independent of the presence or absence of other alternatives. This was tested using the Hausman-McFadden test [25].
- No multicollinearity: Multicollinearity among predictors was checked using variance inflation factor (VIF) values, ensuring all were below the threshold of 5 [26].
- Linearity of logits: The relationship between continuous predictors and the logit of the outcome was confirmed to be linear [27].
- Large sample size: The sample size was sufficiently large to provide reliable estimates of the model parameters [28].

The probability that an observation Y_i belongs to a particular class $c \in \{1, 2, \ldots, C\}$, given the predictor variables $\mathbf{x}_i$, is denoted as $P(Y_i = c|\mathbf{x}_i)$ and is given by

$$P(Y_i = c|\mathbf{x}_i; \alpha_1, \alpha_2, \ldots, \alpha_C, \beta_1, \beta_2, \ldots, \beta_C) = \frac{e^{\alpha_c + \beta_c^T \mathbf{x}_i}}{\sum_{j=1}^{C} e^{\alpha_j + \beta_j^T \mathbf{x}_i}},$$

where α_c denotes the intercept for class c, and β_c represents the vector of regression coefficients for class c.

To assess the performance of the classifier, a confusion matrix is employed to compare the predicted classes to the actual classes from the ground truth data. The accuracy of the model is then calculated by dividing the number of correct predictions by the total number of predictions, thus providing a metric to evaluate the model's performance.

2.3. Random Forest Algorithm

RF is a machine learning model that utilizes an ensemble of decision trees combined with a method known as bootstrap aggregation, or 'bagging'. This approach involves creating multiple subsets of the training data by sampling with replacement from the original dataset [29]. Each subset is used to train a different decision tree, and the forest is constructed by aggregating these trees. This technique often results in higher accuracy compared to a single decision tree model while retaining the interpretability benefits of tree models [30].

In RF, each tree consists of a sequence of binary decisions, or 'nodes', based on a single feature or a combination of features. These nodes create splits in the tree, grouping similar observations together and separating them from others. The algorithm provides measures to rank features according to their importance in predicting the target outcome, selecting those with the highest predictive ability [31].

The RF feature-selection process involves the following steps: training the algorithm on bootstrapped datasets, computing feature importance scores for each tree and selecting features with importance scores above a predefined threshold. The detailed procedure of the RF algorithm is outlined in Algorithm 2.

Algorithm 2 Random Forest (RF)

Input: Training dataset X, number of trees N and feature importance threshold T
Output: Selected features S

Initialize: $S \leftarrow \varnothing$
for $i = 1$ to N **do**
  Sample a bootstrapped dataset X_i from X {Randomly sample with replacement}
  Train a decision tree T_i on X_i
  Compute feature importance scores using T_i
  Update S with features having importance scores above T
end for
return S

The RF feature-selection method is advantageous as it can identify relevant features, reduce dimensionality and improve model interpretability while maintaining high predictive performance.

2.4. Fuzzy Rough Feature-Selection Method

Zadeh's fuzzy set theory [32] extends classical set theory to handle uncertainty and vagueness. In classical set theory, an element either belongs to a set or does not. However, fuzzy set theory allows for partial membership, where elements can belong to a set with varying degrees between 0 and 1. This approach accommodates uncertainty about the boundaries of sets.

Let V be the universe of discourse, which is the set of all possible elements under consideration in a given context. A fuzzy set A is defined as a set of ordered pairs $A = \{(v, \mu_A(v)) : v \in V, \mu_A(v) \in [0,1]\}$, where the membership function $\mu_A : V \to [0,1]$ represents the degree of membership of element v in the fuzzy set A.

Rough set theory aims to handle incomplete or imprecise information by distinguishing between certain and uncertain knowledge [33]. Let U be the universe of objects, which includes all specific objects under analysis in the given context. While V represents all possible elements considered, U focuses on the particular set of objects being studied. Let $R \subseteq U \times U$ be a relation representing the lack of knowledge about elements of U. For a set of objects $X \subseteq U$, $R(x)$ denotes the equivalence class of R determined by element x. The rough set X is composed of the tuple $\langle X^-, X^+ \rangle$, where X^- is the lower approximation and X^+ is the upper approximation, defined as

$$X^- = \bigcup_{x \in U} \{R(x) : R(x) \subseteq X\}$$

$$X^+ = \bigcup_{x \in U} \{R(x) : R(x) \cap X \neq \varnothing\}$$

In the rough set, X^- represents the certain elements of X and X^+ includes both certain and uncertain elements. The boundary region X^0 delineates regions of complete information (positive region) and regions of uncertainty (boundary region): $X^0 = X^+ - X^-$.

The fuzzy rough sets [34] combine fuzzy set theory and rough set theory to handle uncertainty and imprecision in data. Let V be the universe of discourse, and $A \subseteq V$ be a

fuzzy set in V. Let U be a non-empty universe, and R be a similarity relation on U. For $F \in f(U)$, the fuzzy rough set is the pair $\langle R^-(F), R^+(F)\rangle$ of fuzzy sets on U, such that for every $u \in U$

$$R^-(F)(u) = \inf_{y \in U} \max\{1 - R(u,y), F(y)\}$$

$$R^+(F)(u) = \sup_{y \in U} \min\{R(u,y), F(y)\}$$

The fuzzy rough sets feature-selection (FRFS) algorithm identifies relevant features that distinguish between different classes or concepts in a dataset. It uses rough set approximations to evaluate feature contributions and iteratively refines feature selection to balance relevance and redundancy in the selected feature subset. The algorithm considers the dataset's inherent properties to ensure optimal feature selection.

The algorithm selects features that increase the positive region size until it matches the size of the positive region with all features or the required number of features. The concept of indiscernibility is central to locating a positive region. Suppose we have a non-empty set of objects U and a non-empty set of attributes A. The indiscernibility of two objects u_i and u_j based on the sets of attributes in F, where $F \subseteq A$, is given by

$$\text{IND}(F) = \{(u_i, u_j) \in U^2 \mid \forall f \in F, f(u_i) = f(u_j)\}$$

Using indiscernibility, we can identify the partition of U generated by IND(F). This partition is defined as

$$U/\text{IND}(F) = \otimes\{U/\text{IND}(\{f\}) : f \in F\},$$

where $\otimes$ for two sets A and B is represented by

$$A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \varnothing\}$$

Let Q be an equivalence relation over U. The positive region $POS_F(Q)$ can be found using

$$POS_F(Q) = \bigcup_{Y \in U/Q} FY$$

where FY are the rough sets of lower approximations, defined as

$$FY = \{u_i \in U \mid [u_i]_{U/IND(F)} \subseteq Y\}.$$

If the goal is to identify the largest positive region for a specific number of features, denoted as k, the FRFS algorithm calculates the $POS_F(Q)$ for each set of k features and selects the set that maximizes it. This means it chooses the set with the highest number of u_i. Alternatively, if we want to consider all possible combinations of features, from $n = 1$ to $n = n$, FRFS will give us the smallest number of features that maximize $POS_F(Q)$.

The fuzzy rough nearest neighbor (FRNN) classification method can handle fuzzy and uncertain data. It extends the nearest neighbor classification concept by classifying test objects based on their similarity to a specified number of neighbors, denoted by K. The method considers the membership degrees of these neighbors to the class labels when assigning a class label to the test object [35].

The FRNN algorithm computes the distances $\|y - u_j\|$ between an unclassified object y from the testing data and each object u_j from the training data. After calculating these distances, FRNN considers the fuzzy memberships of the k nearest neighbors to determine the fuzzy membership of y to each class c. The aggregation process combines the fuzzy memberships of the neighbors to determine the fuzzy membership of the test object. The

membership degree of a given new sample y in a class c, represented by the k nearest neighbors, is measured as

$$\mu_c(y) = \frac{\sum_{j=1}^{k} \mu_{cj} \left(\frac{1}{\|y-u_j\|^{2/(q-1)}} \right)}{\sum_{j=1}^{k} \left(\frac{1}{\|y-u_j\|^{2/(q-1)}} \right)},$$

where $q \in (1, +\infty)$ is the fuzzy strength parameter and μ_{cj} is the membership of the sample u_j from the training data to the class c among the k nearest neighbors.

To classify the test data point y, we select the class c with the highest fuzzy membership value $\mu_c(y)$. Subsequently, we compare the predicted class labels (associated with fuzzy memberships) to the true class labels (i.e., the original class of the test data point). The accuracy is calculated by determining the correctness of predictions for each data point y.

The algorithm for the FRFS is presented in Algorithm 3.

Algorithm 3 Fuzzy Rough Feature Selection (FRFS)

Input: Dataset D with features $F = \{f_1, f_2, \ldots, f_n\}$, target variable Y
Output: Selected feature subset F_s

Initialize: $F_s \leftarrow \emptyset$
Calculate initial positive region $POS_F(Q)$ for all features F
while $|F_s| < n$ **do**
  for each feature $f_i \in F \setminus F_s$ **do**
    Compute the positive region for feature f_i
  end for
  Find the feature f^* that maximizes the positive region
  Add f^* to F_s
  Update the positive region $POS_{F_s}(Q)$
  if $POS_{F_s}(Q)$ equals $POS_F(Q)$ **then**
    break
  end if
end while
return F_s

3. Overview of the Dataset

This section provides an overview of the dataset used in this study, including the data-collection process, the characteristics of the dataset and the categorization of variables.

3.1. Dataset Description

This dataset is part of a survey administered to 383 randomly selected households in Brazil to collect data about monthly electrical power consumption and the characteristics of the households and occupants. The inputs consist of the building's geographical and structural data and data related to the occupants' social, economic and behavioral aspects. The outputs are low-, medium- and high-load-consumption profiles. Those profiles are formed by clustering the household power consumption with k-means, agglomerative hierarchical clustering (AHC) and self-organizing maps (SOM).

Based on the research conducted by [36], selecting input variables for the dataset depends on expert opinions and initial testing. In this specific case, the selection of each variable results from a study on the factors that affect household electricity consumption.

Various researchers have divided them into different categories to understand better the factors that affect indoor environmental quality. For instance, factors affecting household energy consumption have been categorized into seven groups: climate, building characteristics, social and economic factors, user presence, building service systems, occupant behavior and indoor environmental quality [37]. Additionally, these factors have been

grouped into four categories: external conditions, physical characteristics of the dwelling, appliance and electronics stock, and occupant factors [38]. A simplified classification divides these factors into objective and subjective categories, where objective factors do not depend on the individual's intention, while subjective factors are related to an individual's intention [6].

3.2. Variable Categorization

Machine learning models for energy consumption rely on various features such as outdoor weather conditions, building size and surface area, and dwelling typology [39]. For instance, weather conditions, such as temperature and rainfall, are considered for urban or rural areas where a dwelling is located. Similarly, building characteristics, such as the number of rooms, determine the dwelling typology.

According to a review conducted by ref. [2], heating and air conditioning (HAC) systems, along with domestic hot water (DHW) systems, are some of the major energy-consuming appliances in households. This variation in energy consumption is attributed to differences in thermal comfort expectations, resulting in temperature adjustments, and to different degrees of importance placed on occupant behavior [40].

According to ref. [1], age and family size are important factors linked to energy consumption. Homes with fewer children have more flexibility in their energy-consumption practices compared to homes with children, where parents have less flexibility in their routines. The number of people living in a house is also a significant factor affecting energy consumption, as energy usage tends to increase with the number of occupants [41]. Additionally, ref. [42] suggests that homes with fewer children can perform different practices and consume energy differently, making them more flexible in their energy consumption.

Household disposable income and education level are important factors that influence energy consumption across various energy load profiles [43]. When families have higher incomes, they tend to own more appliances and thus consume more energy. However, a higher income can also lead to investments in modern energy-efficient equipment, thereby decreasing energy consumption.

Each row in the dataset is a twelve-element array consisting of eleven characteristics and one of three cluster types. To provide a clearer understanding of the dataset, Table 2 presents the description of the characteristics, and Table 3 provides the statistical analysis of the numerical variables, including their mean, median, standard deviation, minimum and maximum values.

Table 2. Description of features in the behavior dataset.

Feature	Description
ARE	Area in which the dwelling is located (0-Urban, 1-Countryside)
TYP	Dwelling typology (0-Standard house, 1-Standard apartment, 2-Duplex or triplex)
BED	Number of bedrooms
HAC	Number of heating and air conditioning units in use
DHW	Number of domestic hot water units in use
COE	Usage of electric cooktop and/or electric oven (0-Yes, 1-No)
WPU	Usage of electric water pump (0-Yes, 1-No)
NPE	Number of people living in the dwelling
U18	Number of people under 18 living in the dwelling
PET	Presence of pets, specifically dogs and/or cats (0-Yes, 1-No)
INC	Value of family income (0-Up to R$ 1903.98, 1-From R$ 1903.99 to R$ 4664.68, 2-Above R$ 4664.68)

Table 3. Statistical analysis of numerical variables in the dataset.

Variable	Mean	Std Dev	Min	Median	Max
Number of bedrooms	2.83	1.00	1.00	3.00	8.00
Number of heating and air conditioning units	0.94	1.13	0.00	1.00	5.00
Number of domestic hot water units	1.04	0.93	0.00	1.00	5.00
Number of people	3.16	1.23	1.00	3.00	7.00
Number of people under 18	0.81	0.92	0.00	1.00	5.00

4. The Feature-Selection Process

Generating a model for predictive purposes using data is often challenged by the curse of dimensionality, which can be mitigated by selecting relevant features from the original feature set. This section aims to achieve this by applying specific feature-selection criteria and obtaining a subset of the relevant features.

4.1. Chi-Square Test

In the context of feature selection for a predictive model, the CHI2 test is used to evaluate the independence between each feature and the target variable [44]. Before analyzing the categorical variables, the five numerical features (BED, HAC, DHW, NPE, and U18) were categorized and the categories are presented in Table 4.

Table 4. Categorization of quantitative variables.

Feature	Categories
BED	0—One bedroom, 1—Two bedrooms, 2—Three bedrooms, 3—Four or more bedrooms
HAC	0—Zero HAC, 1—One HAC, 2—Two HACs, 3—Three or more HACs
DHW	0—Zero DHW, 1—One DHW, 2—Two DHWs, 3—Three or more DHWs
NPE	0—One person, 1—Two persons, 2—Three persons, 3—Four or more persons
U18	0—Zero child, 1—One child, 2—Two children, 3—Three or more children

The process involves calculating the *p*-value for each feature to determine its statistical significance in relation to the target variable. If the *p*-value for a feature is greater than 0.05 (at a 95% confidence level), we fail to reject the null hypothesis, indicating that the feature is independent of the target variable. Consequently, that feature is removed from further analysis. We then recalculate the *p*-values for the remaining features and continue this process until all features have *p*-values lower than 0.05. This iterative process ensures that only statistically significant features are retained. The implementation of the CHI2 test was done using the scikit-learn 1.3.0 library in Python 3.9.18 [21].

During the initial test, the k-means and SOM algorithms identified the PET feature with the highest *p*-value. In k-means, the *p*-value was approximately 0.6 (Figure 2a), while in SOM, it was almost equal to 1 (Figure 2c). On the other hand, the AHC algorithm identified the ARE feature with the highest *p*-value, around 0.55 (Figure 2b). Therefore, these features were eliminated as their *p*-values exceeded the 0.05 threshold.

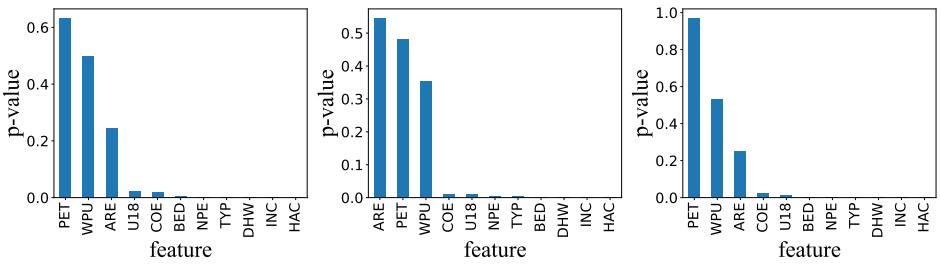
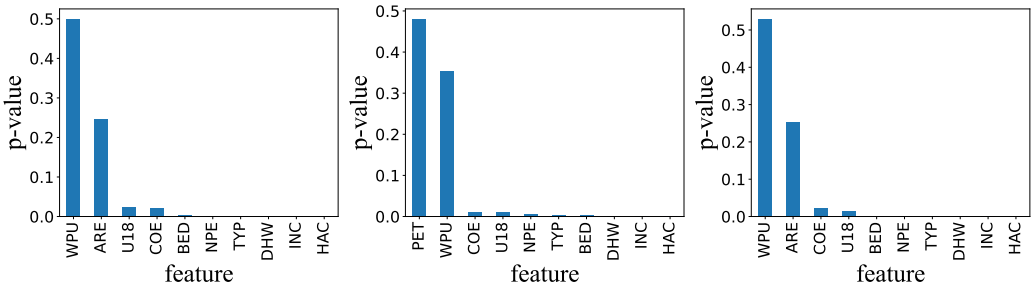


Figure 2. *p*-value of features in the CHI2 test. (**a**) K-means, (**b**) AHC, (**c**) SOM.

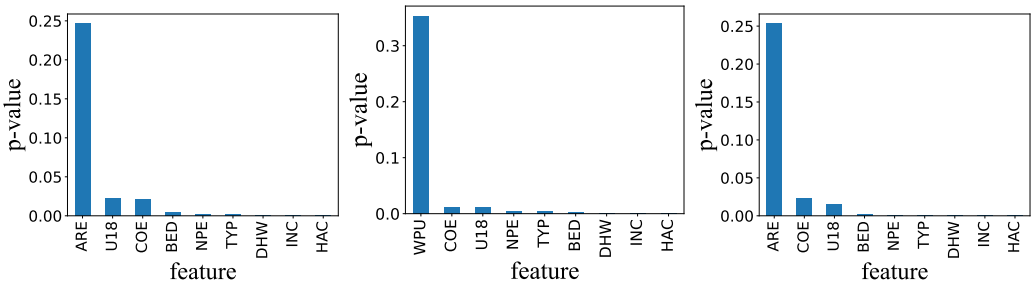
In the second stage of analysis, the WPU feature was removed. The *p*-value was approximately 0.5, slightly less in k-means (Figure 3a) and a little more in SOM (Figure 3c). In the AHC analysis, the PET feature was removed due to its *p*-value being close to 0.5.



(**a**) (**b**) (**c**)

Figure 3. *p*-value of features after removing one feature in the CHI2 test. (**a**) K-means, (**b**) AHC, (**c**) SOM.

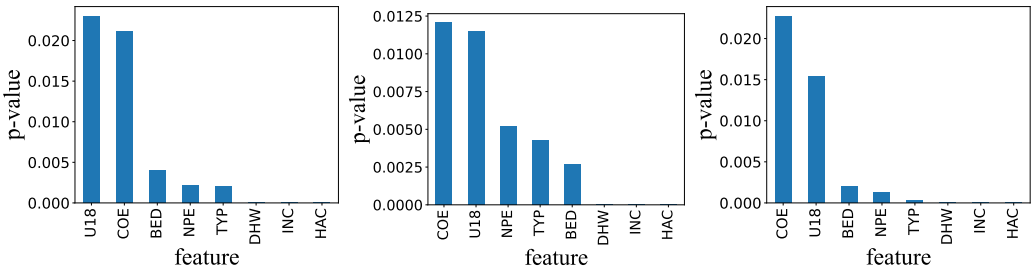
After removing two features, the highest *p*-value continued to be above the 0.05 threshold (Figure 4). The value was closer to 0.25 in all cases: k-means and SOM with the ARE feature, and AHC with WPU.



(**a**) (**b**) (**c**)

Figure 4. *p*-value of features after removing two features in the CHI2 test. (**a**) K-means, (**b**) AHC, (**c**) SOM.

After removing eight features, all *p*-values were below the 0.05 threshold, signaling the end of the process, as shown in Figure 5.



(**a**) (**b**) (**c**)

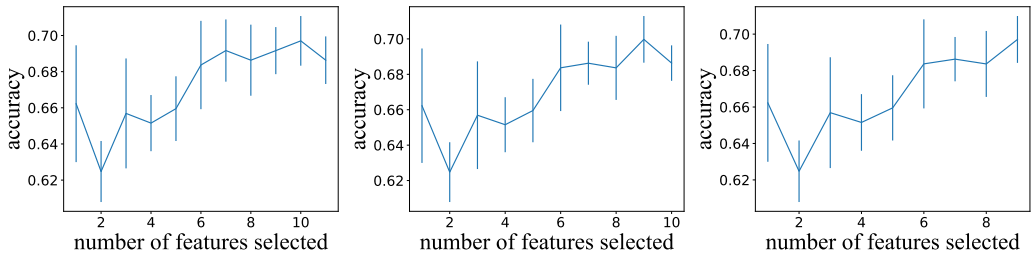
Figure 5. *p*-value of features after removing eight features in the CHI2 test. (**a**) K-means, (**b**) AHC, (**c**) SOM.

After conducting a recursive analysis, the features with a *p*-value higher than the threshold were eliminated until all remaining features had a *p*-value below the threshold. This confirmed that ARE, PET and WPU were independent. As a result, it was concluded that only the following features have a significant relationship with the target variables: HAC, INC, DHW, TYP, NPE, BED, COE and U18.

4.2. Recursive Feature Elimination with Multinomial Logistic Regression

The RFE-MLR determines the optimal number of features required to achieve the best model accuracy. Initially, the algorithm calculates the accuracy of all features and then identifies which feature must be removed to maintain or improve accuracy. This process is repeated until RFE indicates that no more features need to be removed. The algorithms were implemented in Python's scikit-learn library, and stratified k-fold cross-validation was used to enhance performance [21]. To address the potential biases introduced by the initial set of features, we conducted a thorough literature review and consulted with domain experts to identify the most relevant variables for household energy consumption. Additionally, we performed sensitivity analyses to evaluate the impact of including or excluding specific features on the overall model performance. These steps are crucial for ensuring that our feature-selection process remains robust and that the selected features genuinely contribute to understanding energy-consumption patterns.

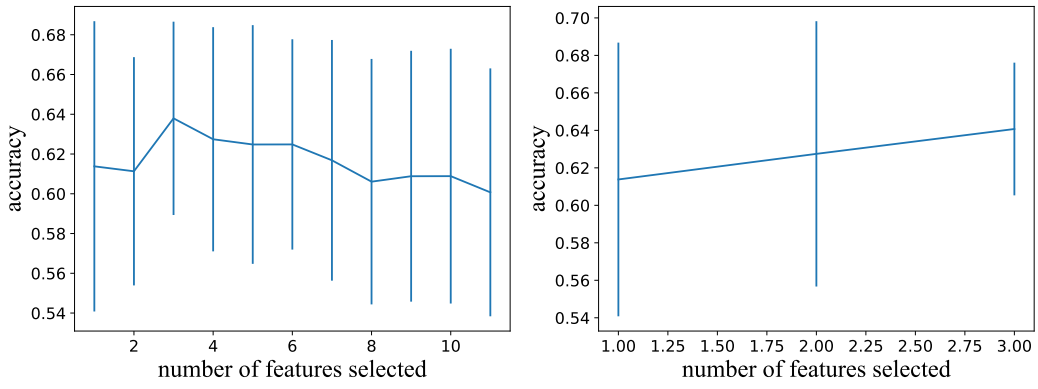
The initial accuracy of RFE-MLR with all eleven features, indicating an optimal number of ten features, was 68.6% using the k-means algorithm (Figure 6a). The algorithm suggested the removal of the U18 feature, after which the accuracy of 68.6% was maintained, and the algorithm then identified an optimal number of nine features (Figure 6b). After removing TYP and running the algorithm for the third time, RFE-MLR showed that the remaining nine characteristics were the optimal number of features (Figure 6c). This means that the k-means algorithm selected the following features as optimal: ARE, HAC, DHW, NPE, WPU, INC, COE, PET and BED, and the model finished with an accuracy of 69.7%.



(**a**) (**b**) (**c**)

Figure 6. Model accuracy vs. the number of features with k-means. (**a**) All features, (**b**) After the 1st removal, (**c**) Process ends.

For the AHC target, we ran the RFE-MLR with all eleven features and obtained an accuracy of 60.1%. The algorithm suggested that only three features are optimal: HAC, DHW and NPE. After removing the non-relevant features, the accuracy increased to 64.1%, as depicted in Figure 7b. Therefore, these three features are considered relevant for the prediction model.



(**a**) (**b**)

Figure 7. Model accuracy vs. number of features with AHC. (**a**) All features, (**b**) Process ends.

According to the RFE-MLR algorithm, only one feature is considered optimal when the target is SOM. This is shown in Figure 8. The model that includes all eleven features has an accuracy of 67.3%. However, when only the HAC feature is used, the accuracy increases to 68.7%.

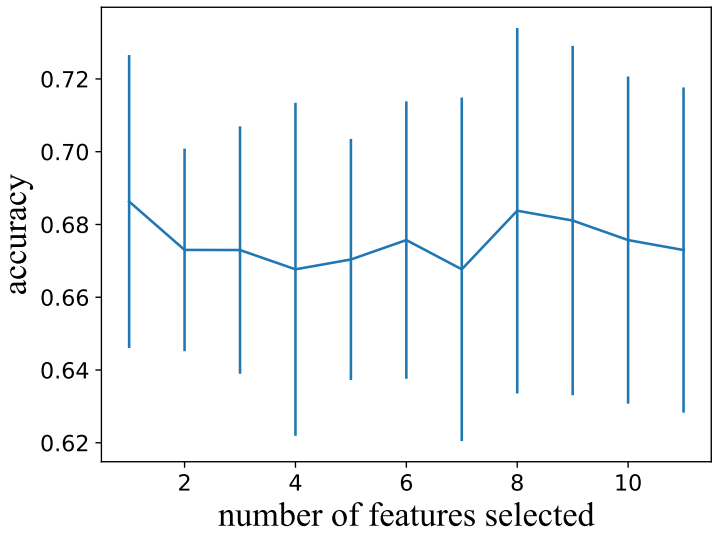


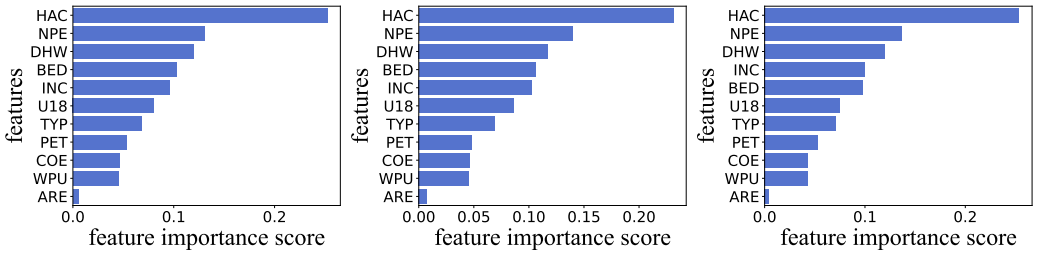
Figure 8. Model accuracy vs. number of features with SOM. All features, process ends.

After analyzing the model accuracy with respect to the choice of factors and using the RFE-MLR algorithm, we found that the algorithm presents different optimal features with each target. Therefore, we have decided to consider the features selected with the k-means algorithm for further analysis. The final accuracy achieved was nearly 70%, and the optimal features selected were ARE, HAC, DHW, NPE, WPU, INC, COE, PET and BED.

4.3. Random Forest Algorithm

Understanding which features have the most significant impact on the predictions of a random forest model is crucial. We calculate the feature importance scores based on how much each feature contributes to the model's accuracy. Features with higher importance scores are more influential. These scores help guide feature selection, where we choose to keep only the most important features for the final model. We then remove each feature with less importance and verify the accuracy. If the accuracy decreases, we stop the process. If not, we repeat the process to simplify the model without sacrificing predictive performance.

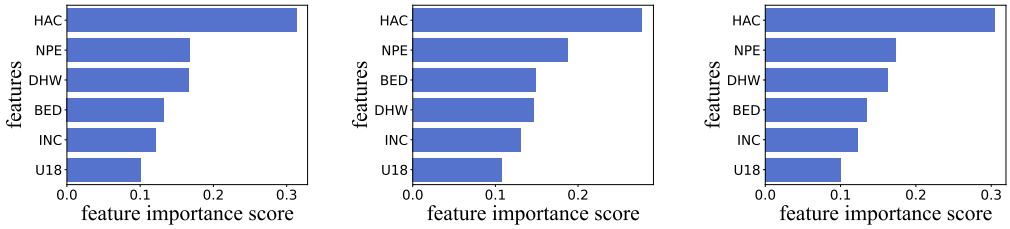
To determine the accuracy of a model, we use a cross-validation algorithm, which helps us understand the significance of the features used. In a random forest (RF) model, the sum of all feature importance scores equals one, or 100%, normalizing each feature's importance within the model's context. We applied this algorithm using the scikit-learn library in Python [21]. In the initial stage, all eleven features produced cluster outputs with the same 70% accuracy. The importance score of the first feature ranking is presented in Figure 9. The analysis reveals that the feature 'ARE' has a relatively small importance score of less than 0.01, representing only 1% of the importance in prediction. On the other hand, the feature 'HAC' has the highest importance score, accounting for almost 25% of the prediction importance.



(**a**) (**b**) (**c**)

Figure 9. Ranking of feature importance with all features in each clustering method as a target. (**a**) K-means, (**b**) AHC, (**c**) SOM.

The RF algorithm first removes the feature 'ARE' and checks for any change in the model accuracy. As there is no change, the process is repeated. The features are dropped individually, based on their importance score in ascending order presented in Figure 9. The RF process stops after dropping ARE, WPU, COE, PET and TYP features, which are not necessarily in this order (Figure 10).



(**a**) (**b**) (**c**)

Figure 10. Ranking of feature importance with selected features in each clustering method as a target. (**a**) K-means, (**b**) AHC, (**c**) SOM.

According to the importance ranking, the feature 'TYP' was removed, and the next feature, 'U18', has an importance score of approximately 10% (as shown in Figure 10). If 'U18' is removed, the model accuracy will decrease to 68% in k-means and AHC, and 69% in SOM. Removing any other features will also negatively impact the model's accuracy. Therefore, the RF process stops here, and the selected features are U18, INC, BED, DHW, NPE and HAC.

4.4. Fuzzy Rough Feature-Selection Method

Determining the ideal number of features can be challenging because there is always a trade-off between the smallest subset and the most accurate modeling [45]. A FRFS can be combined with FRNN for classification to address this challenge. This approach balances both aspects to achieve a more effective feature-selection process.

Combining FRFS with the FRNN algorithm provides a comprehensive framework for handling uncertainty and imprecision in feature-selection and feature-classification tasks. This integration allows for more robust and accurate modeling of complex real-world data, where traditional methods may fall short due to their inability to handle uncertainty effectively. These algorithms are available in the fuzzy-rough-learn Python library [46].

In a random sample of behavioral data (Table 5), we compared the positive region selecting the features $\{f_2, f_3\}$ and $\{f_4, f_5\}$. Here, x_i represents the objects ($x_i \in U$), f represents the attributes ($f \in A$) and p represents the k-means consumption clusters. Table 6 shows the resulting set for the partition $U/IND(F)$ when considering the features $\{f_2, f_3\}$.

Table 5. Random sample of the behavioral dataset.

U	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	p
x_1	1	0	2	1	1	0	1	2	0	0	4	2
x_2	0	0	2	0	0	0	0	3	0	0	4	1
x_3	0	1	3	1	2	1	1	2	0	0	2	2
x_4	0	1	1	1	1	1	1	1	0	1	2	1
x_5	0	0	3	0	1	0	1	3	1	0	4	1
x_6	0	0	2	0	1	1	1	1	0	1	0	1
x_7	0	1	3	2	2	0	1	5	1	1	4	2
x_8	0	0	3	4	2	0	1	3	1	1	4	3

The objects x_1, x_2 and x_6 are indiscernible in relation to the attributes $\{f_2, f_3\}$, as they all have $f_2 = 0$ and $f_3 = 2$. Similarly, x_3 and x_7 cannot be distinguished from each other as they both have $f_2 = 1$ and $f_3 = 3$, and x_5 and x_8 cannot be distinguished from each other as they both have $f_2 = 0$ and $f_3 = 3$. Therefore, the resulting set for the partition $U/IND(F)$ is shown in Table 6.

Table 6. Partition of U based on f_2 and f_3.

$U/IND(F)$	f_2	f_3
$\{x_1, x_2, x_6\}$	0	2
$\{x_3, x_7\}$	1	3
$\{x_5, x_8\}$	0	3
$\{x_4\}$	1	1

Given a set of clustered objects Q, if U/Q is defined as $\{\{x_1,x_3,x_7\}, \{x_2,x_4,x_5,x_6\}, \{x_8\}\}$, the elementary sets presented in $U/IND(f_2, f_3)$, which are also contained in X (where X is a subset of U/Q), are $\{x_3, x_7\}$ and $\{x_4\}$. Therefore, the values of FX and $POS_F(Q)$ for $F = \{f_2, f_3\}$ and $Q = \{p\}$ are

$$FX = \{\{x_3, x_7\}, \{x_4\}\}$$

$$POS_{\{f_2,f_3\}}(Q) = \{x_3, x_4, x_7\}$$

Regarding $\{f_4, f_5\}$, it seems that x_1 and x_4 are indiscernible because both have $f_4 = 1$ and $f_5 = 1$, while x_5 and x_6 both have $f_4 = 0$ and $f_5 = 1$. The resulting set for the partition $U/IND(F)$ is shown in Table 7.

Table 7. Partition of U based on f_4 and f_5.

$U/IND(F)$	f_4	f_5
$\{x_1, x_4\}$	1	1
$\{x_2\}$	0	0
$\{x_3\}$	1	2
$\{x_5, x_6\}$	0	1
$\{x_7\}$	2	2
$\{x_8\}$	4	2

The sets contained in X, which is a subset of U/Q, and presented in $U/IND(f_4, f_5)$ are: $\{x_2\}$, $\{x_3\}$, $\{x_5, x_6\}$, $\{x_7\}$ and $\{x_8\}$. Therefore, $POS_F(Q)$ and FX, where $F = \{f_4, f_5\}$ and $Q = \{p\}$, are:

$$FX = \{\{x_2\}, \{x_3\}, \{x_5, x_6\}, \{x_7\}, \{x_8\}\}$$

$$POS_{\{f_4,f_5\}}(Q) = \{x_2, x_3, x_5, x_6, x_7, x_8\}$$

When $POS_{\{f_4,f_5\}}(Q)$ surpasses $POS_{\{f_2,f_3\}}(Q)$, the combination of features f_4 and f_5 becomes more important than f_2 and f_3.

In the analyzed dataset, FRFS identified the features that increased the positive region for each $n = k$, where k ranged from 2 to 10. The selected features are listed in Table 8.

Table 8. Features selected with maximum $POS_F(Q)$.

n	Selected Features
2	TYP/DHW
3	ARE/TYP/DHW
4	ARE/TYP/BED/DHW
5	ARE/TYP/BED/HAC/DHW
6	ARE/TYP/BED/HAC/DHW/COE
7	ARE/TYP/BED/HAC/DHW/COE/WPU
8	ARE/TYP/BED/HAC/DHW/COE/WPU/NPE
9	ARE/TYP/BED/HAC/DHW/COE/WPU/NPE/U18
10	ARE/TYP/BED/HAC/DHW/COE/WPU/NPE/U18/PET

In this approach, FRNN is combined with FRFS to calculate the model accuracy of each feature group (as shown in Table 8). The n features that achieve the highest accuracy are then selected. If two feature groups have the same accuracy, the group with the lowest n value is chosen.

To maximize accuracy, the FRNN algorithm was executed using 75–25% train and test groups, with k neighbors in $\{3, 5, 7, 9, 11\}$. Each value in Table 9 represents the average accuracy obtained from multiple runs to ensure robustness and reliability of the results.

Table 9. FRNN model accuracy for n features selected.

n	K-Means Accuracy	AHC Accuracy	SOM Accuracy
2	0.6489	0.5957	0.6596
3	0.6489	0.5957	0.6596
4	0.6277	0.5957	0.6595
5	0.6809	0.7021	0.7128
6	0.6596	0.6383	0.6596
7	0.6596	0.6277	0.6809
8	0.6809	0.6383	0.6383
9	0.6596	0.6277	0.6914
10	0.6489	0.6064	0.6489

After applying the FRNN algorithm combined with FRFS, it was determined that the optimal number of features for maximizing accuracy was $n = 5$. Specifically, the selected features were ARE, TYP, BED, HAC, and DHW.

5. Analyzing the Selected Features

In this section, we conduct a trend analysis, assessing the impact of selected features on household energy load profiles and identifying out-profiled households—those that deviate from the typical patterns of their assigned consumption cluster and exhibit behaviors characteristic of a different, typically lower-consumption profile.

5.1. Feature-Selection Overview

The feature selection provides insights into which features the model relies on, but it does not explain why a feature is important. Domain knowledge is crucial for understanding the context and significance of certain features. Through Table 10, we can analyze the results obtained from the feature-selection techniques simultaneously.

Table 10. Features and the methods by which they were selected.

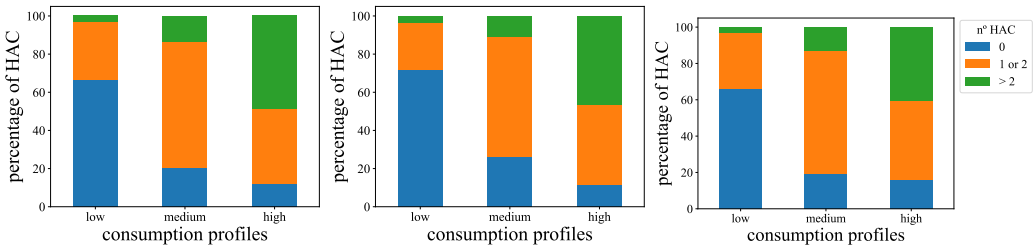
Feature	Method
ARE	RFE-MLR/FRFS-FRNN
TYP	CHI2/FRFS-FRNN
BED	CHI2/RFE-MLR/RF/FRFS-FRNN
HAC	CHI2/RFE-MLR/RF/FRFS-FRNN
DHW	CHI2/RFE-MLR/RF/FRFS-FRNN
COE	CHI2/RFE-MLR
WPU	RFE-MLR
NPE	CHI2/RFE-MLR/RF
U18	CHI2/RF
PET	RFE-MLR
INC	CHI2/RFE-MLR/RF

While feature-selection techniques shed light on which features are influential in the model, they do not inherently provide the rationale behind them. Thus, a deep understanding is essential to comprehend the context and significance of these features. For that analysis, we focus on features common to at least three selection methods.

Thus, we will analyze the number of people living in a dwelling (NPE), the number of heating and air conditioning units (HAC), the number of hot water units (DHW), the number of bedrooms (BED), and the value of family income (INC).

5.2. Heating and Air Conditioning Equipment

The analysis of HAC usage reveals distinct patterns across various clustering methods, as depicted in Figure 11. This pattern underscores the significant influence of HAC equipment on energy-consumption profiles.



(**a**) (**b**) (**c**)

Figure 11. Cumulative bar chart of the HAC percentage in each consumption profile. (**a**) K-means, (**b**) AHC, (**c**) SOM.

In the low-consumption profile, most dwellings do not use HAC. In k-means, 66.42% of dwellings fall into this category, a behavior maintained across other clustering algorithms: 71.74% in AHC and 66.37% in SOM.

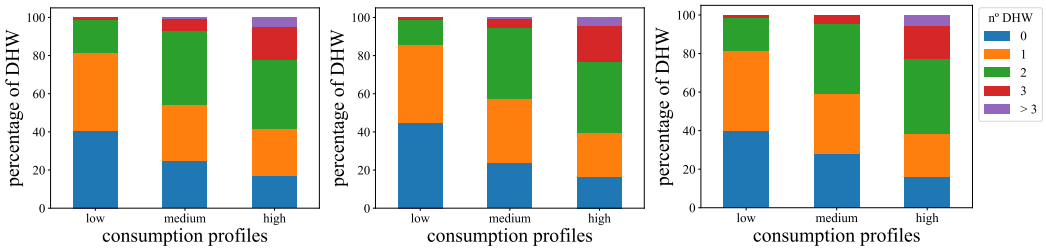
Conversely, the medium-consumption profile is characterized by a substantial increase in HAC presence. The proportion of dwellings without HAC significantly drops to around 20% in k-means and SOM and to 25% in AHC. This shift suggests an escalating dependence on HAC as consumption levels rise.

The trend becomes more pronounced in the high-consumption profile, where the percentage of dwellings without HAC diminishes further. Notably, the proportion of households using more than two HAC units increases markedly in this category, jumping from 13.39% to 48.78% in k-means, 10.96% to 46.51% in AHC, and 12.90% to 40.35% in SOM.

The decreasing trend of dwellings without HAC in the medium- and high-consumption profiles indicates a significant shift towards increased HAC usage. Consequently, households within these profiles that do not use HAC can be considered out-profiled, deviating from the typical pattern of their respective consumption categories.

5.3. Domestic Hot Water Equipment

The use of DHW equipment reveals a clear trend where higher-consumption profiles are associated with an increased presence of DHW units. This correlation is presented in Figure 12.



(**a**) (**b**) (**c**)

Figure 12. Cumulative bar chart of the DHW percentage in each consumption profile. (**a**) K-means, (**b**) AHC, (**c**) SOM.

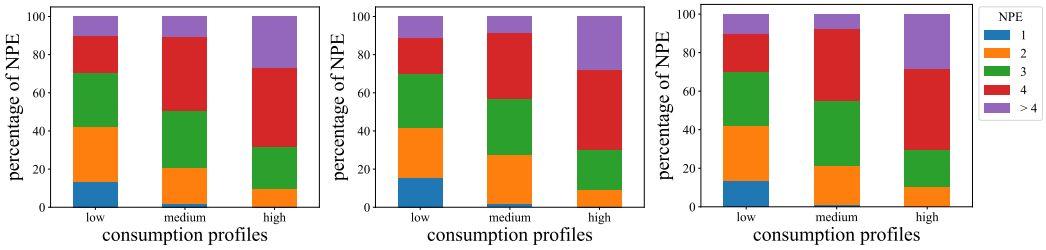
In the low-consumption profile, a significant proportion of dwellings, nearly 80%, have none or only one DHW unit. This figure declines to around 60% in the medium profile and drops to about 40% in the high-consumption profile, indicating a strong correlation between higher energy consumption and the number of DHW units. This pattern is consistent across the k-means, AHC, and SOM clustering methods.

The medium-consumption profile predominantly consists of households with at least one DHW unit, while the high-consumption profile is characterized by dwellings with two or more DHW units. This delineation suggests that the presence of multiple DHW units is a significant indicator of higher energy-consumption levels.

Consequently, households in the medium profile without any DHW unit can be considered out-profiled, deviating from the general trend of this consumption category. Similarly, in the high-consumption profile, dwellings with a maximum of one DHW unit are also deemed out-profiled, given their lower than expected DHW count for this high-energy-consumption category.

5.4. Number of People in Household

It is universally acknowledged that the number of people living in a household (NPE) impacts energy consumption, as evidenced in the analysis across three clustering algorithms (Figure 13).



(**a**) (**b**) (**c**)

Figure 13. Cumulative bar chart of the NPE percentage in each consumption profile. (**a**) K-means, (**b**) AHC, (**c**) SOM.

In the low-consumption profile, approximately 70% of dwellings house up to three individuals. This proportion significantly decreases in the medium- and high-consumption profiles. Notably, dwellings with a single occupant are almost exclusively found in the low-consumption profile, with their presence in the medium and high profiles being minimal or non-existent.

Dwellings with two occupants also constitute a minority in the medium profile, accounting for around 20%, and this figure diminishes to nearly 10% in the high-consumption profile. In contrast, the high-consumption profile predominantly comprises households with four or more occupants.

Therefore, it can be inferred that dwellings with up to two occupants in the medium-consumption profile and those with up to three occupants in the high-consumption profile are outliers or 'out-profiled' households, deviating from the typical occupancy patterns of their respective consumption categories.

5.5. Number of Bedrooms

The number of bedrooms across different household energy-consumption profiles provides insightful correlations. In Figure 14, we categorize dwellings based on the number of bedrooms, including a category for more than four bedrooms, given the low proportion of such dwellings in the dataset.

Dwellings with up to three bedrooms are predominantly found in the low-consumption profile, accounting for nearly 90% of households. This trend sharply decreases in the medium- and high-consumption profiles, where the prevalence of one and two-bedroom dwellings is significantly lower. Single-bedroom dwellings are virtually absent in the high-consumption profile, and two-bedroom dwellings constitute a small fraction (around 4.88% in k-means, 4.65% in AHC and 5.26% in SOM).

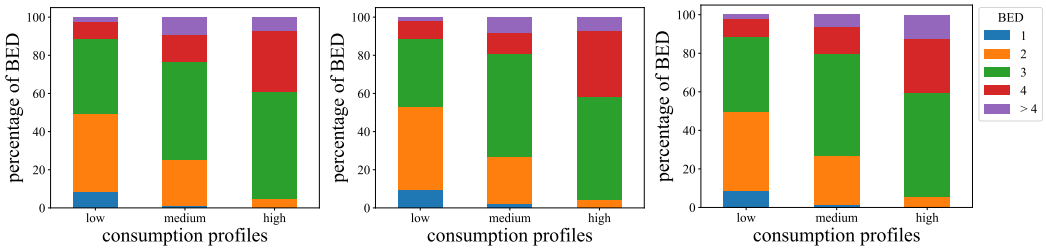


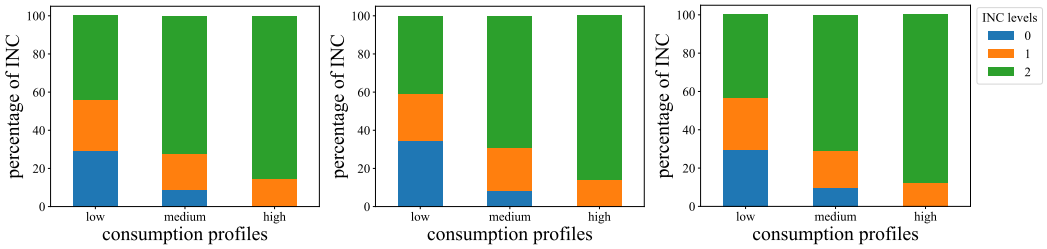
Figure 14. Cumulative bar chart of the BED percentage in each consumption profile. (**a**) K-means, (**b**) AHC, (**c**) SOM.

Conversely, the proportion of dwellings with three or more bedrooms increases with the consumption profiles. This category represents almost half of the dwellings in the low-consumption profile, about three-quarters in the medium-consumption profile and close to 95% in the high-consumption profile, regardless of the clustering method used.

This analysis reveals that dwellings with up to two bedrooms in the medium- and high-consumption profiles deviate from the common pattern, thus qualifying as out-profiled households.

5.6. Household Income Levels

The analysis of household energy consumption concerning income levels is presented in Figure 15, showcasing the distribution of various income levels across different consumption profiles.



(**a**) (**b**) (**c**)

Figure 15. Cumulative bar chart of the INC levels percentage in each consumption profile. (**a**) K-means, (**b**) AHC, (**c**) SOM.

In the medium-consumption profile, dwellings at the lowest income level comprise only about 9%, which reduces to zero in the high-consumption profile. A similar decreasing trend is observed for the second income level, representing around 19% in the medium profile (22% in the AHC profile) and never exceeding 15% in the high profile, irrespective of the clustering method.

Considering that most households in both the medium- and high-consumption profiles fall into the highest income category, it is reasonable to infer that dwellings within these profiles' first and second income levels are outliers or 'out-profiled' households. These findings emphasize the influence of income on household energy consumption and the divergence of lower-income households from the typical consumption patterns of their respective profiles.

6. Conclusions and Future Work

This study has provided significant advancements in the analysis of household energy consumption by examining data from 373 dwellings in Brazil using four distinct feature-selection methods: the chi-square (CHI2) independence test, recursive feature elimination with multinomial logistic regression (RFE-MLR), random forest (RF), and fuzzy rough feature selection combined with fuzzy rough nearest neighbor (FRFS-FRNN). These methods were instrumental in identifying key features associated with energy consumption, offering a deeper understanding of consumer behaviors and aiding in identifying out-profiled consumers. The findings from this research can guide future studies in energy consumption, including evaluating new techniques and exploring additional data and features across different regions.

From the initial eleven features included in the survey, five were retained after the feature-selection analysis. Subsequent individual analyses confirmed that these features significantly impact household energy consumption across low-, medium-, and high-load profiles.

Key findings of this study include the identification of out-profiled dwellings, which are characterized by the absence of heating and air conditioning (HAC) equipment in medium- or high-consumption profiles, the absence of domestic hot water (DHW) equipment in the medium profile, and the presence of only one DHW unit in the high profile. Additionally, dwellings with up to two people in the medium profile and up to three people in the high profile, as well as those with up to two bedrooms in medium or high profiles, were considered out-profiled. Households with an income of up to R$ 4664.68 in medium or high profiles also fell into this category.

While this study presents notable advancements, it is not without limitations. Firstly, the research is geographically constrained to Brazilian households, which might limit the generalization of findings to other regions with different climatic, cultural, and socio-economic conditions. Secondly, the effectiveness of the feature-selection methods is contingent on the initial set of features in the dataset. The exclusion of certain variables or the presence of unobserved confounding factors might have influenced the study's outcomes.

Despite these limitations, the study opens several directions for future research. One potential path is the development of dynamic energy tariff models based on consumer behavior, as suggested by our findings, which can lead to more equitable billing systems

and promote energy conservation. Additionally, applying these techniques in different geographical and socio-economic contexts can validate and possibly enhance the generalization of our findings. Further research could also explore the integration of additional variables, such as real-time energy-usage data and more granular socio-economic indicators, to deepen the understanding of household energy behavior.

In conclusion, this study contributes significantly to the field of energy management by offering a nuanced understanding of household energy behavior through advanced feature-selection techniques. Our findings facilitate the development of tailored strategies for sustainable energy consumption and highlight the potential for more equitable energy management practices. Building on these insights, addressing the identified limitations and exploring new methodologies will foster a more comprehensive understanding of household energy dynamics.

Author Contributions: Conceptualization, L.H., C.C. and F.P.L.; data curation, L.H., C.C. and F.P.L.; formal analysis, L.H., C.C. and F.P.L.; investigation, L.H., C.C. and F.P.L.; methodology, L.H., C.C. and F.P.L.; writing—original draft preparation, L.H., C.C. and F.P.L.; writing—review and editing, L.H., C.C. and F.P.L. All authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: Research was partially financed by Portuguese funds through the CMAT—Research Centre of Mathematics of University of Minho, Portugal, within projects UIDB/00013/2020 (https://doi.org/10.54499/UIDB/00013/2020, accessed on 1 June 2024) and UIDP/00013/2020 (https://doi.org/10.54499/UIDP/00013/2020, accessed on 1 June 2024) (Lucas Henriques, Cecilia Castro).

Data Availability Statement: The codes and datasets used in this study are available upon request.

Acknowledgments: The authors would also like to thank the Editors and reviewers for their constructive comments, which led to improvements in the presentation of the article. The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Csutora, M.; Zsoka, A.; Harangozo, G. The Grounded Survey–An integrative mixed method for scrutinizing household energy behavior. *Ecol. Econ.* **2021**, *182*, 106907. [CrossRef]
2. Wei, Y.; Zhang, X.; Shi, Y.; Xia, L.; Pan, S.; Wu, J.; Han, M.; Zhao, X. A review of data-driven approaches for prediction and classification of building energy consumption. *Renew. Sustain. Energy Rev.* **2018**, *82*, 1027–1047. [CrossRef]
3. Caldera, M.; Hussain, A.; Romano, S.; Re, V. Energy-consumption pattern-detecting technique for household appliances for smart home platform. *Energies* **2023**, *16*, 824. [CrossRef]
4. Szymańska, E.J.; Kubacka, M.; Polaszczyk, J. Households' energy transformation in the face of the energy crisis. *Energies* **2023**, *16*, 466. [CrossRef]
5. Ma, P.; Cui, S.; Chen, M.; Zhou, S.; Wang, K. Review of family-level short-term load forecasting and its application in household energy management system. *Energies* **2023**, *16*, 5809. [CrossRef]
6. Zhou, K.; Yang, S. Understanding household energy consumption behavior: The contribution of energy big data analytics. *Renew. Sustain. Energy Rev.* **2016**, *56*, 810–819. [CrossRef]
7. Copiaco, A.; Himeur, Y.; Amira, A.; Mansoor, W.; Fadli, F.; Atalla, S.; Sohail, S. An innovative deep anomaly detection of building energy consumption using energy time-series images. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105775. [CrossRef]
8. Borgato, N.; Prataviera, E.; Bordignon, S.; Garay-Martinez, R.; Zarrella, A. A data-driven model for the analysis of energy consumption in buildings. In Proceedings of the 53rd AiCARR International Conference "From NZEB to ZEB: The Buildings of the Next Decades for a Healthy and Sustainable Future", Milan, Italy, 12–14 March 2024; Volume 523, p. 02002.
9. Karananos, A.; Dimara, A.; Arvanitis, K.; Timplalexis, C.; Krinidis, S.; Tzovaras, D. Energy Consumption Patterns of Residential Users: A Study in Greece. In *Computer Vision Systems*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 639–650.
10. Zheng, P.; Zhou, H.; Liu, J.; Nakanishi, Y. Interpretable building energy consumption forecasting using spectral clustering algorithm and temporal fusion transformers architecture. *Appl. Energy* **2023**, *349*, 121607. [CrossRef]
11. Shaikh, A.; Nazir, A.; Khalique, N.; Shah, A.; Adhikari, N. A new approach to seasonal energy consumption forecasting using temporal convolutional networks. *Results Eng.* **2023**, *19*, 101296. [CrossRef]
12. Hora, C.; Dan, F.; Bendea, G.; Secui, C. Residential short-term load forecasting during atypical consumption behavior. *Energies* **2022**, *15*, 291. [CrossRef]

13. Ramesh, G.; Logeshwaran, J.; Kiruthiga, T.; Lloret, J. Prediction of energy production level in large PV plants through AUTO-encoder based neural-network (AUTO-NN) with restricted Boltzmann feature extraction. *Future Internet* **2023**, *15*, 46. [CrossRef]
14. He, P.; Zhou, Y.; Qin, X. A Survey on Energy-Aware Security Mechanisms for the Internet of Things. *Future Internet* **2024**, *16*, 128. [CrossRef]
15. Jeon, H.; Oh, S. Hybrid-recursive feature elimination for efficient feature selection. *Appl. Sci.* **2020**, *10*, 3211. [CrossRef]
16. Sheikhpour, R.; Sarram, M.A.; Gharaghani, S.; Chahooki, M.A.Z. A survey on semi-supervised feature-selection methods. *Pattern Recognit.* **2017**, *64*, 141–158. [CrossRef]
17. Zhang, L.; Wen, J. A systematic feature selection procedure for short-term data-driven building energy forecasting model development. *Energy Build.* **2019**, *183*, 428–442. [CrossRef]
18. Miao, J.; Niu, L. A survey on feature selection. *Procedia Comput. Sci.* **2016**, *91*, 919–926. [CrossRef]
19. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A.W.; O'Sullivan, J.M. A review of feature-selection methods for machine learning-based disease risk prediction. *Front. Bioinform.* **2022**, *2*, 927312. [CrossRef] [PubMed]
20. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *Acm Comput. Surv.* **2017**, *50*, 1–45. [CrossRef]
21. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
22. Bahassine, S.; Madani, A.; Al-Sarem, M.; Kissi, M. Feature selection using an improved Chi-square for Arabic text classification. *J. King Saud-Univ.-Comput. Inf. Sci.* **2020**, *32*, 225–231. [CrossRef]
23. Misra, P.; Yadav, A.S. Improving the classification accuracy using recursive feature elimination with cross-validation. *Int. J. Emerg. Technol* **2020**, *11*, 659–665.
24. Jeune, W.; Francelino, M.R.; Souza, E.d.; Fernandes Filho, E.I.; Rocha, G.C. Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti. *Rev. Bras. Ciênc. Solo* **2018**, *42*, e0170133. [CrossRef]
25. Hausman, J.A.; McFadden, D. Specification tests for the multinomial logit model. *Econom. J. Econom. Soc.* **1984**, 1219–1240. [CrossRef]
26. Kutner, M.H.; Nachtsheim, C.; Neter, J.; Li, W. *Applied Linear Regression Models*; McGraw-Hill: New York, USA, 2004.
27. Agresti, A. *Foundations of Linear and Generalized Linear Models*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
28. Cochran, W.G. *Sampling Techniques*; John Wiley & Sons: New York, NY, USA, 1977.
29. Karasu, S.; Altan, A. Recognition Model for Solar Radiation Time Series Based on Random Forest with Feature Selection Approach. In Proceedings of the 2019 11th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, 28–30 November 2019; pp. 8–11.
30. Speiser, J.L.; Miller, M.E.; Tooze, J.; Ip, E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* **2019**, *134*, 93–101. [CrossRef]
31. Janitza, S.; Tutz, G.; Boulesteix, A.L. Random forest for ordinal responses: Prediction and variable selection. *Comput. Stat. Data Anal.* **2016**, *96*, 57–73. [CrossRef]
32. Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353. [CrossRef]
33. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [CrossRef]
34. Dubois, D.; Prade, H. Rough Fuzzy Sets and Fuzzy Rough Sets. *Int. J. Gen. Syst.* **1990**, *17*, 191–209. [CrossRef]
35. Jensen, R.; Cornelis, C. Fuzzy-rough nearest neighbour classification and prediction. *Theor. Comput. Sci.* **2011**, *412*, 5871–5884. [CrossRef]
36. Torabi, M.; Hashemi, S.; Saybani, M.R.; Shamshirband, S.; Mosavi, A. A Hybrid clustering and classification technique for forecasting short-term energy consumption. *Environ. Prog. Sustain. Energy* **2019**, *38*, 66–76. [CrossRef]
37. Yu, Z.; Fung, B.C.; Haghighat, F.; Yoshino, H.; Morofsky, E. A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy Build.* **2011**, *43*, 1409–1417. [CrossRef]
38. Kavousian, A.; Rajagopal, R.; Fischer, M. Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy* **2013**, *55*, 184–194. [CrossRef]
39. Amasyali, K.; El-Gohary, N.M. A review of data-driven building energy consumption prediction studies. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1192–1205. [CrossRef]
40. Yildiz, B.; Bilbao, J.I.; Dore, J.; Sproul, A.B. Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Appl. Energy* **2017**, *208*, 402–427. [CrossRef]
41. Guo, Z.; Zhou, K.; Zhang, C.; Lu, X.; Chen, W.; Yang, S. Residential electricity consumption behavior: Influencing factors, related theories and intervention strategies. *Renew. Sustain. Energy Rev.* **2018**, *81*, 399–412. [CrossRef]
42. Malatesta, T.; Breadsell, J.K. Identifying Home System of Practices for Energy Use with K-Means Clustering Techniques. *Sustainability* **2022**, *14*, 9017. [CrossRef]
43. Mi, L.; Xu, T.; Sun, Y.; Yang, H.; Wang, B.; Gan, X.; Qiao, L. Promoting differentiated energy savings: Analysis of the psychological motivation of households with different energy consumption levels. *Energy* **2021**, *218*, 119563. [CrossRef]
44. Miola, A.C.; Miot, H.A. Comparing categorical variables in clinical and experimental studies. *J. Vasc. Bras.* **2022**, *21*, e20210225. [CrossRef]

45. Jensen, R.; Shen, Q. Fuzzy-rough sets assisted attribute selection. *IEEE Trans. Fuzzy Syst.* **2007**, *15*, 73–89. [CrossRef]
46. Lenz, O.U.; Peralta, D.; Cornelis, C. Fuzzy-rough-learn 0.1: A Python library for machine learning with fuzzy rough sets. In Proceedings of the International Joint Conference on Rough Sets, Havana, Cuba, 9 June–3 July 2020; pp. 491–499.