*sustainability*

MDPI

Article

# Benchmarking Evaluation of Building Energy Consumption Based on Data Mining

**Thomas Wu, Bo Wang, Dongdong Zhang *, Ziwei Zhao and Hongyu Zhu ***

School of Electrical Engineering, Guangxi University, Nanning 530004, China
* Correspondence: dongdongzhang@gxu.edu.cn (D.Z.); hongyuzhu@st.gxu.edu.cn (H.Z.)

**Abstract:** University building energy consumption is an important proportion of the total energy consumption of society. In order to work out the problem of poor practicability of the existing benchmarking management method of campus building energy consumption, this study proposes an evaluation model of campus building energy consumption benchmarking management. By analyzing several types of feature data of buildings, this study uses random forest method to determine the building features that have outstanding contributions to building energy consumption intensity and building classification, and uses the K-means method to reclassify buildings based on the building features obtained after screening, to obtain a building category that is more in line with the actual use situation and to solve the problem that the existing building classification is not in line with the reality. Compared with the original classification method, the new classification method showed significant improvement in many indexes, among which DBI decreased by 60.8% and CH increased by 3.73 times. Finally, the quart lines of buildings in the category of new buildings are calculated to obtain the low energy consumption line, medium energy consumption line and high energy consumption line of buildings, so as to improve the accuracy and practicability of energy consumption line classification.

**Keywords:** building energy consumption; benchmarking; data mining; random forest model; factor analysis; K-means cluster

check for updates

**Citation:** Wu, T.; Wang, B.; Zhang, D.; Zhao, Z.; Zhu, H. Benchmarking Evaluation of Building Energy Consumption Based on Data Mining. *Sustainability* **2023**, *15*, 5211. https://doi.org/10.3390/su15065211

Academic Editor: Dušan Katunský

Received: 12 February 2023
Revised: 10 March 2023
Accepted: 13 March 2023
Published: 15 March 2023

## 1. Introduction

The energy crisis and climate change have resulted in higher requirements for urban energy conservation and emission reduction [1]. Large-public buildings, which have high rates of energy consumption and great energy-saving potential, are an indispensable part of reducing urban carbon emissions. In 2019, the United States used 16% of end-use energy in the residential sector, and 12% in the commercial sector, implying that energy efficiency in buildings has made and can still make a substantial impact on energy conservation efforts as a whole [2]. In China, campus buildings are important examples of large-public buildings, accounting for 19% of the total buildings constructed from 2001 to 2020 [3]. With the growth of the campus scale, the energy consumption of campus buildings is also increasing, which is of great significance regarding the progress towards carbon neutrality [4].

The current study of building energy consumption mainly includes the following aspects. The first method establishes a benchmark model of building energy consumption through simulation calculation. Common building dynamic simulation tools such as Department of Energy 2(DoE-2) and EnergyPlus consider various specific characteristics of buildings for dynamic simulation [5,6]. Zhang Xu [7] also calculated the all-round energy consumption of typical residential buildings in Shanghai by using the Bin method and calculated the influence of factors such as the heat transfer coefficient of exterior walls and windows, design parameters of indoor air conditioning and energy efficiency ratio on building energy consumption by changing some parameters. The dynamic energy consumption simulation method considers all kinds of indoor and outdoor perturbations more carefully, and the results are more accurate; however, the dynamic simulation project is very complicated. With the help of the computer, it is convenient to dynamically calculate the cooling and heating load of buildings under the action of changing outdoor parameters.

In the second method, multiple fitting regression technology is used to establish the benchmark model of building energy consumption, which is mainly used to forecast building energy consumption through regression analysis to obtain the benchmarking of architecture energy consumption [8–10]. However, most of the analysis and studies on building energy intensity obtained by linear regression model lack accuracy, and there may be multicollinearity among explanatory variables. The Energy Star project in the United States also sets up a regression model and a scoring system to let users know the energy consumption level of the building. However, this method does not consider other characteristics such as equipment and operation plan, which will also affect the energy-saving performance of the building [11], and requires a lot of data. Input data may not be available in the simulation, which further increases its complexity. In recent years, the study of building energy consumption by machine learning method has attracted much attention, and some studies have adopted the data-driven machine learning method to examine problems. For example, random forest algorithm (RF) [12], support vector (SVM) algorithm [13] and artificial neural network algorithm (ANN) [14] are all adopted to deal with the research on building energy conservation. Bu, Shao and Wang used the support vector machine (SVR) model to estimate the energy consumption of hotel buildings [15]. In their study, RBF kernel function was selected as the kernel function of SVR, and the accuracy of model prediction was improved by optimizing kernel parameters. In another study, Kim, Jung and Kang used a residential energy consumption prediction model based on neural network [16]. During the modeling process, South Korean residential building information and user characteristics were taken into account. Their study discusses the share of influencing parameters, that is, the number of external walls, housing orientation, housing size, years of residence, number of family members and occupation of the head of household, on energy estimates. Their results show that the neural network model has high accuracy in predicting energy consumption. These machine learning approaches are currently a hot topic for data-driven models.

The third method mainly uses other methods to study building energy consumption quota, which is achieved by evaluating the energy performance of buildings relative to similar buildings of the same type or geographical area. Among them, the "Government Energy Efficiency Best Practice Project" in the United Kingdom classifies the same type of building twice and evaluates the level of energy consumption and energy cost, respectively [17]. The building energy consumption evaluation standard VDI3807 in Germany evaluates the energy consumption level of the building by comparing the overall energy consumption of the building with the reference value. Using the upper quartile as the target energy consumption level of this type of building [18], Santamouris [19] et al. classified the energy consumption of campus buildings in Greece, involving electricity and heating energy, and used the equal frequency method to calculate the cumulative frequency of the statistical samples as 20%, 40%, 60%, 80% and 100%. Building energy consumption is defined as the five types of building energy consumption benchmarking namely A, B, C, D and E; Hernandez [20] et al. used the quartile method to study the energy consumption of primary schools in Ireland; they obtained the cumulative distribution curve of energy consumption of 88 primary schools, and found that the average energy consumption index was 96 kW·h/m$^2$, which was taken as the benchmarking of building energy consumption, and obtained the first quartile of 65 kW· h/m$^2$. As a building energy consumption reference line. Salah Vaisi and Pouya Varmazyari et al. [21] developed a top-down energy benchmark based on actual energy consumption within large government office buildings. Through a survey of 26 office buildings in cold climates, four general benchmark levels were developed, including "Best practices," "good practices," "Benchmark," and "bad practices." Paola Marrone, Paola Gori and Francesco Asdrubali et al. [22] conducted a cluster analysis of the stock of school buildings in Italy, aiming to provide a method to identify the best energy retrofit interventions from a cost-effective perspective, and relate them to the specific characteristics of educational architecture. It also provides a model that can be developed into a useful tool for public administration to set priorities in planning for the new energy retrofit of existing school buildings. Research in this approach focuses on comparing energy performance between buildings and identifying potential areas for improvement.

There are plenty of buildings in the campus of colleges and universities. According to the benchmarking line measurement method of many buildings, it is important to study the classification of buildings [3,22]. Now, the study of campus energy consumption benchmarking is few, which cannot be used in actual energy conservation research. According to their main functions, buildings on university campuses can be divided into two categories: residential buildings and public buildings. Since the living buildings in colleges and universities account for an important proportion of staff dormitories and family buildings, and these two categories are generally not calculated for campus energy consumption quota [23], the research object of this paper is public buildings in colleges and universities, and the goal is to obtain the energy consumption reference line of college buildings. According to the existing studies [24–29], the EUI of a building is not only limited by the building function, but also affected by the number of floors, number of degree days and other factors. Therefore, the same types of buildings also have a big difference in energy use mode due to the different building characteristics. The traditional method also puts this type of building in the scope of transformation when defining energy consumption benchmarking, which is inconsistent with our original intention of setting building energy consumption benchmark. To solve this problem, it is proposed to reclassify the buildings. At the same time, there is the phenomenon of building function combination in the university campus, that is, a building has more than one function at the same time. Due to the different proportions and types of each functional area, the composite functional buildings have different energy use rules, therefore, they have different amounts of energy saving potential. Under this premise, traditional building classification methods cannot classify buildings according to specific functions, and the obtained building energy consumption benchmarking cannot meet the demand for energy conservation benchmarking.

The purpose of this study is to obtain the overall energy consumption benchmark of different buildings. Benchmarking refers to the energy consumption index of a building and the limit value of the energy consumption level of a building within a certain limited range, which is used to guide the energy conservation work of a building. This paper selects building energy intensity (EUI), which can best represent the overall level of building energy consumption. To solve the problems of inaccurate and poor practicability of the existing campus architecture energy consumption benchmarking, this paper puts forward a new evaluation method of campus building energy consumption benchmarking, which mainly includes the following aspects:

(1) The classification in the original campus building classification model is inconsistent with the reality. Influenced by the number of floors, degrees and other factors, and based on the data-driven idea, through the random forest mining of important features affecting building energy consumption and classification, with the building EUI and the original building classification label as targets, we build a random forest model and determine the importance ranking of each building feature.

(2) Factor analysis is adopted to reduce the dimension of the studied features according to the importance of architectural features, and several common factors affecting architectural classification are extracted to reduce subsequent clustering errors. Then, K-means clustering method is used to cluster the extracted common factors of architectural features, and a new architectural classification is obtained. Compared with the original classification method, the new classification method has significant improvement in many indexes.

(3) On the basis of the above methods, based on the study of building energy consumption factors by the second type of benchmark research method, we do not directly obtain specific energy consumption lines, but use the statistical method in the third type of energy consumption benchmark research to compare the target building with the same

type of building, so as to delimit the low, medium and high energy consumption lines, and to make the energy consumption benchmark more practical. A more accurate and practical reference line for energy consumption of campus buildings is obtained by calculating the quartile line among the buildings of the same type.

## 2. Benchmarking Assessment of Building Energy Consumption

### *2.1. New Campus Building Benchmarking Assessment Method*

Buildings with different uses differ greatly in terms of building feature data, such as weekly working hours, external structure, internal number of elevators and internal number of computers [30]. Therefore, one potential solution is to divide buildings into new building classification according to relevant characteristics to solve the problem at hand, namely that the original building classification does not meet the reality; thus, the actual energy consumption reference line will be able to be obtained on the basis of the new classification. This paper presents a new method for measuring the reference line of building energy consumption.

### *2.2. Random Forest Algorithm to Determine the Importance of Features*

The random forest algorithm is an integrated learning algorithm based on decision tree and random subspace theory. The basic idea is to construct several base learners with different levels of performance, and combine the prediction results of base learners through certain strategies [31]. Random forest improves the prediction accuracy without significant increase in the amount of computation. Random forest is not sensitive to multivariate collinearity, and the results are relatively robust to missing data and unbalanced data. It can accurately predict the effects of up to thousands of explanatory variables, and is known as one of the best algorithms at present.

Random forest is a classification model designed to build forests in a random manner. The forest consists of a number of decision trees, each of which is unrelated to the other. After the random forest model is obtained, when the new sample enters the random forest, each decision tree in the random forest is judged separately. After the results are obtained, the category or one of the most voted categories is usually used as the final model output. In order to avoid the inaccuracy and collinearity caused by the linear method, this paper chooses the random forest model to solve the problem. Since both EUI and building classification are correlated with building characteristics in building energy consumption data [32], the random forest model can be used to determine which important features affect building EUI and building classification.

The random forest model structure diagram of Figure 1. Ref. [33] is established in this paper. Firstly, a new training sample set is generated by randomly sampling and repeated random sampling of N samples, and then N classification trees are generated according to the self-sample set to form a random forest. For each node, m features are randomly selected, and the decision of each node in the decision tree is determined based on these features. According to these m features, the optimal splitting mode was calculated [34]. Each tree will grow intact without pruning, and then all N trees will be averaged to reduce the overall model variance [35].

In this paper, due to the excessive variables studied, it is not conducive to further analyze the problem; therefore, it is necessary to rank the importance of building features that affect building EUI and building classification, so as to provide a reference basis for removing unimportant features to obtain decisive features. According to the contribution of each feature obtained by the random forest algorithm, the importance ranking of each feature to the building feature is obtained, and the features that have an important impact on the research target are screened out, which also provides a reference for the subsequent dimension reduction.

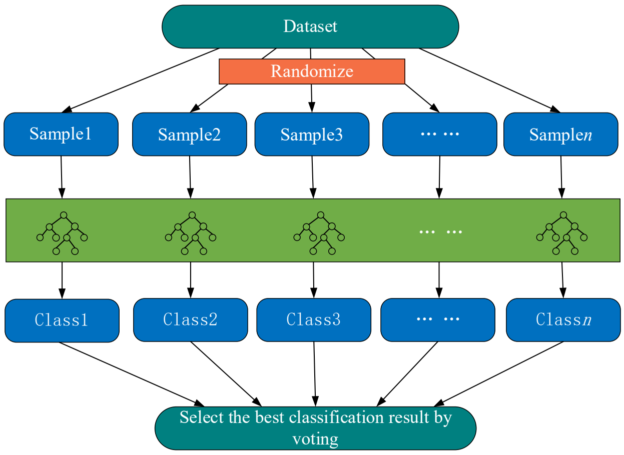

**Figure 1.** Random forest structure diagram.

### *2.3. Factor Analysis Is Used to Extract Common Factors of Building Features*

After obtaining the feature importance ranking that affects building EUI and building classification, we understand the influence degree of each building feature studied on building EUI and classification. However, clustering algorithm cannot be used to divide clustering clusters at this time, for the following reasons: First, the building data without extracting common factors are clustered according to K-means, which may lead to the buildings in the building cluster only having mathematical connection but no practical connection. Secondly, there may be correlations among the building features involved in the clustering. If there is correlation, the results will also be biased, because it increases the weight of a certain type of feature virtually, which is difficult to avoid when screening features. In order to solve the above problems, the method of factor analysis is used to extract the common factor. In this paper, principal component analysis method is selected to extract the factor. This is carried out not only to reduce the original building data dimension, but also to complete its common extraction, and discover the actual relationship between different building features. Factor analysis mainly includes the following steps.

In order to solve the above problems, the method of factor analysis is used to extract the common factor. In this paper, the principal component analysis method is selected to extract the relevant factor. It not only completes the dimensionality reduction in the original data but also carries out the generic extraction to discover the actual connection between different features. Factor analysis mainly includes the following steps:

(1) Correlation investigation among variables. Factor analysis requires a strong correlation between original variables. Commonly used tests include the Kaiser–Meyer–Olkin (KMO) test for correlation coefficient and partial correlation coefficient between variables and Bartlett spherical test for independence. The calculation formula of KMO is as follows:

$$\varepsilon_{\text{KMO}} = \frac{\sum\limits_{i}\sum\limits_{j(i\neq j)} r_{ij}^2}{\sum\limits_{i}\sum\limits_{j(i\neq j)} r_{ij}^2 + \sum\limits_{i}\sum\limits_{j(i\neq j)} s_{ij}^2} \tag{1}$$

where $s_{jj}$ and $r_{ij}$ are the correlation coefficient and partial correlation coefficient of variable $X$, respectively. The closer the value of the KMO test is to 1, the more suitable for factor analysis, whereas a value less than 0.5 is considered not suitable for factor analysis. Additionally, Bartlett's null hypothesis is the identity matrix of the correlation coefficient matrix. The statistic is obtained from the determinant of this matrix. According to whether its significance level less than 0.05, to determine whether it is suitable for factor analysis.

(2) Principal component extraction. By solving the eigen value ($\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p$) of the correlation coefficient matrix of the original variable and the corresponding

orthonormal eigenvectors $(u_1, u_2, \ldots, u_p)$, and selecting the previous eigenvector with eigenvalues greater than 1 as the principal component of the original variable for analysis.

(3) Factor load matrix calculation. The factor load matrix determined by each principal component is defined as shown in Equation (2), $\left(\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p\right)$ are the eigenvalue of the correlation coefficient matrix, $(u_1, u_2, \ldots, u_p)$ are the corresponding eigenvector.

$$A = \begin{bmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1p}\sqrt{\lambda_p} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2p}\sqrt{\lambda_p} \\ \cdots & & & \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \cdots & u_{pp}\sqrt{\lambda_p} \end{bmatrix} \tag{2}$$

(4) Factor rotation. According to the elementary load matrix, the contribution rate of each common factor is calculated, and m principal factors are selected. By rotating the extracted factor load matrix, the matrix $B = A_m T$ (where $A_m$ is the front m column of $A$, and $T$ is the orthogonal matrix) is obtained, and the factor model is constructed.

(5) Calculate the factor score. Using the regression method, the common factor F and variables $(X_1, X_2, X_3..)$, make the regression, establish regression equation and then substitute variable values into the regression equation. The relationship between the factor and the original variable is shown in Equation (3). $a$ is each element of the load matrix, which is essentially the correlation coefficient between the common factor and the original variable. The factor score is obtained according to Equation (4).

$$\begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_\mathrm{m} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \cdots \\ F_\mathrm{m} \end{bmatrix} \tag{3}$$

$$\begin{bmatrix} F_1 \\ F_2 \\ \cdots \\ F_\mathrm{m} \end{bmatrix} = \begin{bmatrix} \mathrm{b}_{11} & \mathrm{b}_{12} & \cdots & \mathrm{b}_{1\mathrm{p}} \\ \mathrm{b}_{21} & \mathrm{b}_{22} & \cdots & \mathrm{b}_{2\mathrm{p}} \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{b}_{\mathrm{m}1} & \mathrm{b}_{\mathrm{m}2} & \cdots & \mathrm{b}_{\mathrm{mp}} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_\mathrm{m} \end{bmatrix} \tag{4}$$

### *2.4. Building Evaluation Cluster Analysis Based on K-Means*

Cluster analysis is a statistical method that creates a group of objects or clusters in which the objects in one cluster are very similar and the objects in different clusters are very different. For numerical data, a cluster is usually a set of numbers in which each number is closer to its average, that is, the average of all the numbers in the clustering, than to the average of any other cluster. The goal is to minimize the sum of the difference between each number and its closest average. This algorithm is based on the clustering of centers or prototypes. The K-means algorithm is a typical partitioning-based clustering algorithm [36]. Its basic idea is that a given sample set is divided into K clusters according to the distance between the given samples. The samples in each cluster have levels of similarity, whereas the similarity of different clusters is low. Supposing that the data of the cluster are divided into $(C_1, C_2, \ldots, C_k)$, our goal is the to minimize the squared error $E$:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu\|_2^2 \tag{5}$$

where $\mu_i$ is the mean vector of cluster $C_i$, also known as the center of mass:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{6}$$

The schematic diagram of the K-means clustering algorithm in this paper is shown in Figure 2. The specific steps of the algorithm are as follows:

(1) Randomly set K feature space points as the initial building clustering center;
(2) Calculate the distance between the points corresponding to other buildings and K centers, and select the nearest cluster center point as the marker category for the unknown points;
(3) Place the points corresponding to each building against the labeled cluster center, and recalculate the new center point of each cluster;
(4) If the calculated new center point is the same as the original center point, the algorithm will be terminated; otherwise, return to the second step.

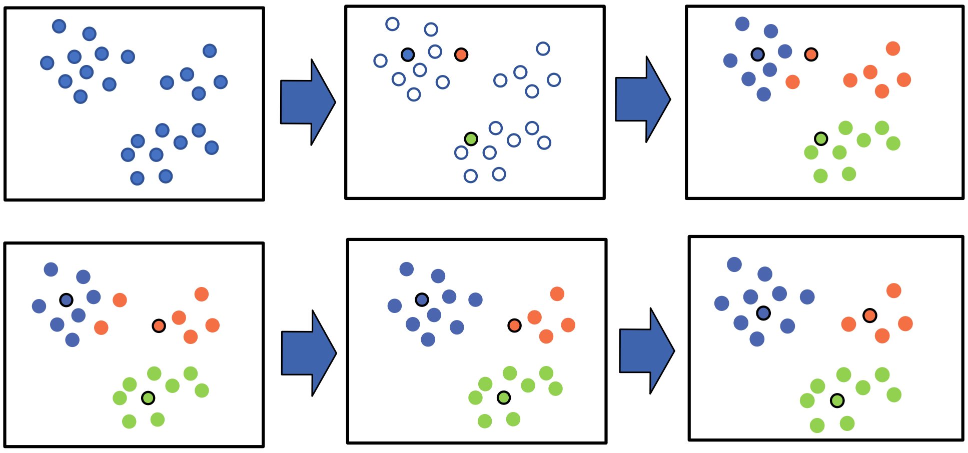

**Figure 2.** Schematic diagram of K-means clustering.

We use the factor analysis method to obtain the common factors of building features. On this basis, the K-means clustering method is used to divide the building set into several clustering clusters, and a new building classification is obtained, which should be able to better meet the needs of demarcating the new benchmarking in line with the actual situation.

### *2.5. Evaluation of Clustering Effect*

Clustering involves randomness; thus, it is difficult to judge the effect of clustering. The clustering effectiveness index can help us measure the clustering effect after clustering a group of data, and then make a judgment according to the actual significance of the data. We usually judge the effect of clustering by the compactness of cluster and the degree of separation between cluster. Compactness is a measure of whether the sample points in a cluster are compact enough, such as the average distance from the cluster center. The degree of separation is a measure of whether the sample is far enough away from other clusters. We choose two indexes to judge the clustering results. Respectively, they are Davide-Bouldin index (DBI) and Calinski-Harabasz index (CH), where a smaller DBI means smaller intra-class distance and larger inter-class distance. The formula is:

$$\mathrm{DBI} = \frac{1}{k}\sum_{i=1}^{k}\max\left(\frac{avg(C_i)+avg(C_j)}{d_c(\mu_i,\mu_j)}\right) \tag{7}$$

where $k$ represents how many clusters there are in the cluster, $\mu_i$ represents the center point of the $i$th cluster and $\mathrm{avg}(C_i)$ represents the average distance between all data in the $i$th cluster and the center point of the $i$th cluster. $d_c(\mu_i,\mu_j)$ represents the distance between the center of the $i$th cluster and the center of the $j$th cluster.

The larger the CH index is, the closer the class itself is and the more dispersed among classes, which means better clustering results. The formula is:

$$\mathrm{CH} = \frac{tr(B_k)}{k-1} / \frac{tr(W_k)}{N-k} \tag{8}$$

where $k$ represents the number of cluster categories and $N$ represents the total number of data. $tr(B_k)$ is the variance between classes, $tr(W_k)$ is the variance within classes.

## 3. Methods

### *3.1. Data Preprocessing*

The database used in this study comes from Commercial Building Energy Consumption Survey (CBECS) [37] data and includes a sample of buildings across most of the United States, representing commercial buildings in all 50 states and the District of Columbia. CBECS is a national sample survey that collects information on the commercial building stock in the United States, including energy-related building characteristics and energy use data.

Database data collection consists of two phases. The first stage is the building survey, which collects building features and energy usage data (annual consumption and cost) from respondents through interviews or web-based questionnaires. The second stage is the Energy Supplier Survey (ESS), which is a follow-up survey to the energy supplier of the building that responded to the first stage. Suppliers of electricity, natural gas, heating oil (including fuel oil, kerosene and diesel) and district heating (steam or hot water) provide monthly data on the energy use of each building [38]. For most buildings, these files contain information such as building dimensions, year of construction, type of energy used, energy consumption and expenditure. Since the database contained incomplete building information, data filtering was applied to overcome technical limitations. The screened data include partially missing data, missing data that are not applicable after interpolation and identical data. At the same time, because the object of the study is campus buildings, non-campus buildings are removed and only campus buildings are retained according to the owner category and building type in the sample features [39].

These more than 200 building samples were obtained after data filtering. In this paper, some types of buildings are selected for discussion because of the content to be studied. Since the database contains too many features with a low correlation to building energy consumption, features should be selected. According to experience, features unrelated to university buildings should be excluded. Meanwhile, the distinguishing variables are selected based on experience, as shown in Table 1.

**Table 1.** Building features and serial numbers.

| No. | Feature |
|---|---|
| 1 | One activity in building |
| 2 | Area |
| 3 | Building shape |
| 4 | Exterior glass ratio |
| 5 | Number of floors |
| 6 | Year of construction |
| 7 | Month used |
| 8 | Working hours |
| 9 | Office equipment |
| 10 | CDD |
| 11 | HDD |
| 12 | Heat percent |

**Table 1.** *Cont.*

| No. | Feature |
|---|---|
| 13 | Cool percent |
| 14 | Number of elevators |
| 15 | Number of computers |

After obtaining the subsets of the building data studied, in order to eliminate the dimensional influence among the indicators, a data standardization processing is needed to solve the comparability among the data indicators. The numerical features in the original data were treated with Z-Score standardization, as shown in Equation (9).

$$x_{new} = \frac{x-\mu}{\sigma} \tag{9}$$

where $\mu$ is the mean and $\sigma$ is the standard deviation of the sample data.

In addition, such features are also included in the building features to illustrate whether the building is a multifunctional composite building and the occupied area proportion of various functions, as shown in Table 2.

**Table 2.** Cases of functional characteristics of buildings.

| Feature | Sample | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Single function | No | No | Yes |
| First function | Office | Education | Office |
| First percent | 60 | 60 | 100 |
| Second function | Education | Auditorium | None |
| Second percent | 40 | 40 | 0 |

The type of the building sample is the type of the maximum functional area. For example, the first function of sample 1 is an office building, then the label of the building sample is an office building. However, the above table can only obtain the area proportion of each functional area but not the actual energy consumption of each functional area. Therefore, the maximum functional area in the database statistics may not represent the actual function of the building. This phenomenon also exists in practice, that is, the original label of the building is not a good classification of the building. Therefore, the building samples need to be reclassified.

### *3.2. Screen Building Features through Random Forest*

In order to obtain a new building classification, we first built a random forest model according to the data set, in which 70% of the original data were divided a into training set and 30% into a test set, the target value was set as the EUI of the building and the power consumption, area, building shape, exterior glass ratio, number of floors, building year, service months and weekly working hours were set as inputs. The importance ranking of building features obtained through the construction of random forest model is shown in Figure 3.

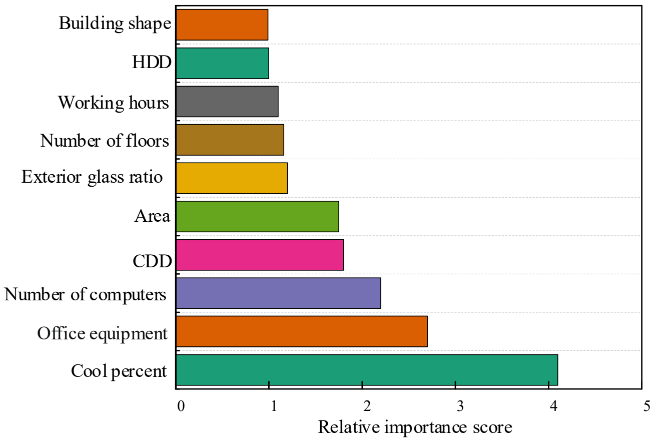

**Figure 3.** The target value is the building feature importance ranking of EUI.

The EUI factor is an important indicator of building energy consumption. However, since all the characteristics of the building cannot be collected during the building information collection stage, the collected characteristic information may only show part of the building information. Therefore, only referring to EUI factor analysis may lead to the final result being only responsible for building EUI while ignoring other factors. There was only a mathematical connection between the results and no consideration of the actual working form of the building; thus, we considered taking other factors into account. During the analysis stage of this paper, we believe that the original label of a building may not be completely consistent with reality, but we also believe that the original label of a building is related to the initial design purpose of the building; therefore, it has a certain reference value for the classification of buildings. Buildings with the same label may not be classified into one category, but many of them have a certain degree of similarity. Therefore, this paper considers this factor, sets the target value as the original label of the building, and sets the input as power consumption, area, building shape, exterior glass ratio, floor number, building year, service months and weekly working hours. The importance ranking of building features obtained by the random forest model is shown in Figure 4.

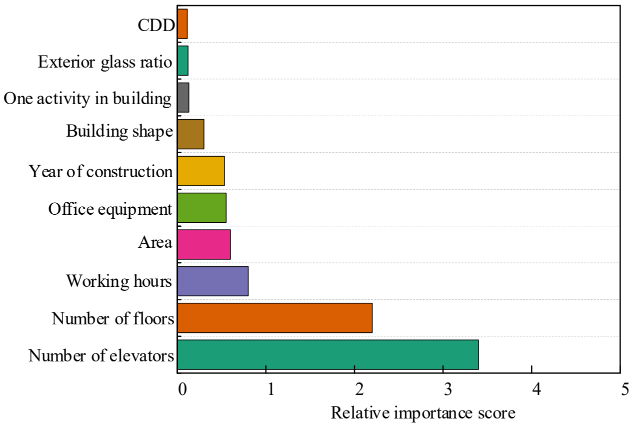

**Figure 4.** The target value is the building feature importance ranking of the original classification of the building.

After combining the two feature importance rankings, we obtained seven features that have important contributions to building EUI and the original label of buildings: the number of floors, the proportion of exterior glass, the area, the load of office equipment, the number of computers, the power consumption and the building year, while ignoring the features that only have an important impact on a certain factor to make our consideration more practical and specific. The research of this paper continues on this basis.

### *3.3. Reduces the Dimension of the Remaining Building Features by Factor Extraction*

We have obtained building features that have a great impact on building EUI and the original classification of buildings, as shown in Table 3. However, at this time, the clustering algorithm cannot be used to divide clustering clusters, because large errors may be generated at this time. Therefore, factor analysis was carried out on the data. Principal component analysis was used for factor extraction, and variance maximum rotation method was used for rotation.

**Table 3.** Building features after screening.

| No. | Feature |
|---|---|
| 1 | Office equipment |
| 2 | CDD |
| 3 | Area |
| 4 | Exterior glass ratio |
| 5 | Number of floors |
| 6 | Working hours |
| 7 | Number of computers |

The common factor was extracted for the remaining features, the KMO test statistic was greater than 0.5 and the Bartlett sphericity test $p$-value < 0.05, consistent with the premise of factor extraction. According to the lithotripsy diagram, the optimal number of common factors is 2, and the factor load matrix after rotation and the two principal components whose initial eigenvalue is greater than 1 is extracted to obtain the principal components F1 and F2. Therefore, they can be identified as the two global principal components to be extracted. At the same time, the score of each factor can be calculated according to the factor score coefficient and the standardized value of the original variable, to extract the principal component of the obtained factor.

### *3.4. Cluster the Extracted Common Factors by K-Means*

After the above steps are completed, the principal component extracted from factor analysis is taken as the clustering feature, and the building data in the case are clustered using the K-means algorithm.

The K value in the k-means algorithm is given in advance, and the initial clustering K value in most kinds of literature is randomly drawn up according to experience, but the size of the K value is difficult to estimate in general, and different K values often lead to widely different clustering results. In this paper, the silhouette coefficient method and sum variance value (SSE) is used to determine the initial K value. Among them, the closer the silhouette coefficient is to 1, the better the clustering effect is, and the larger the SSE value is, the larger the error is. However, in practice, the "elbow" of the SSE is often taken as the optimal case, that is, the SSE changes the most before and after this point. The silhouette coefficients under different K values were obtained as shown in Figure 5, and the SSEs under different K values are shown in Figure 6.

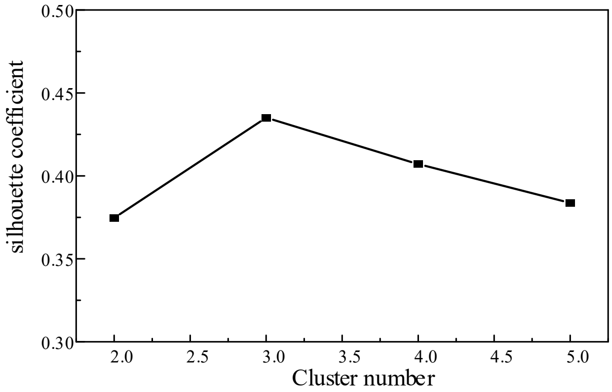

**Figure 5.** Clustering silhouette coefficient diagram.

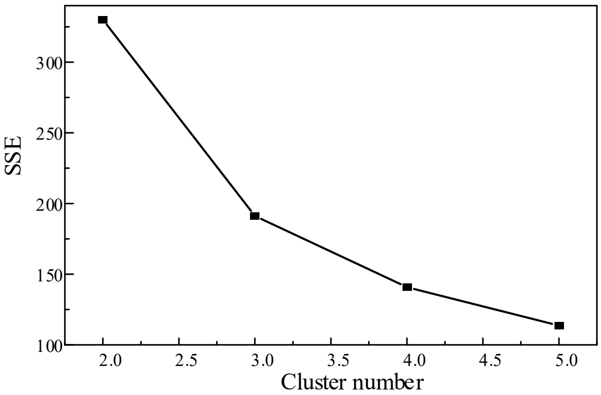

**Figure 6.** Clustering SSE value diagram.

We know that the silhouette coefficient reaches its maximum when K is 3, which is the “elbow” of the SSE value graph; thus, the optimal K value is 3. Therefore, the optimal cluster number is determined to be 3. The clustering results are shown in the following figure, and the clustering results are shown in Figure 7. According to the figure, the teaching building and office building in the case can be better divided into three categories. This may meet the classification needs of buildings.

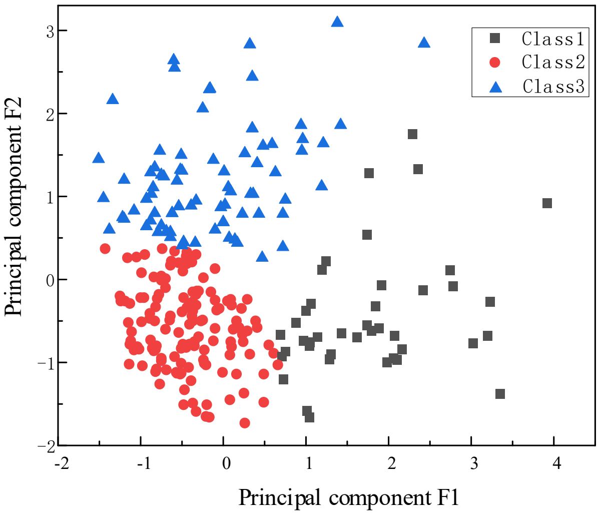

**Figure 7.** Results clustering.

### 3.5. New Building Energy Benchmarking

We compared buildings within each building category in order to obtain energy benchmarks for that building category. We use the quartile method to measure the building energy consumption benchmark for the new building classification, as shown in Figure 8. At this point, the median line defined by the quartile line can represent most levels of the building. Since our building classification is more realistic, the energy consumption baseline we obtained is obviously more realistic than the baseline drawn on the basis of the original building classification. After clustering by the K-means method, the original mixed-function buildings are also divided, and the reclassification diagram is shown in Figure 9.

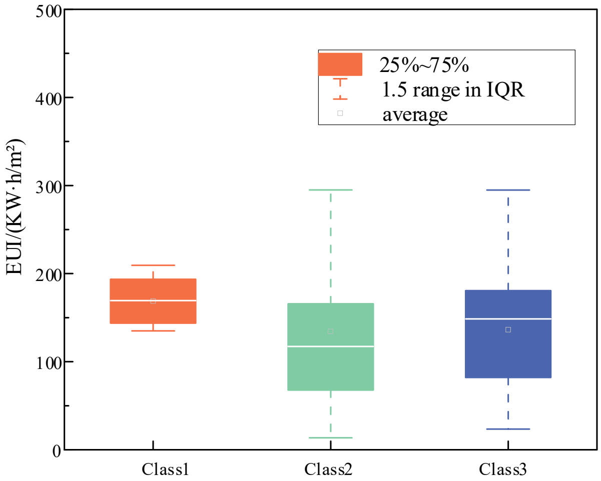

**Figure 8.** The quartile diagram obtained with the new method.

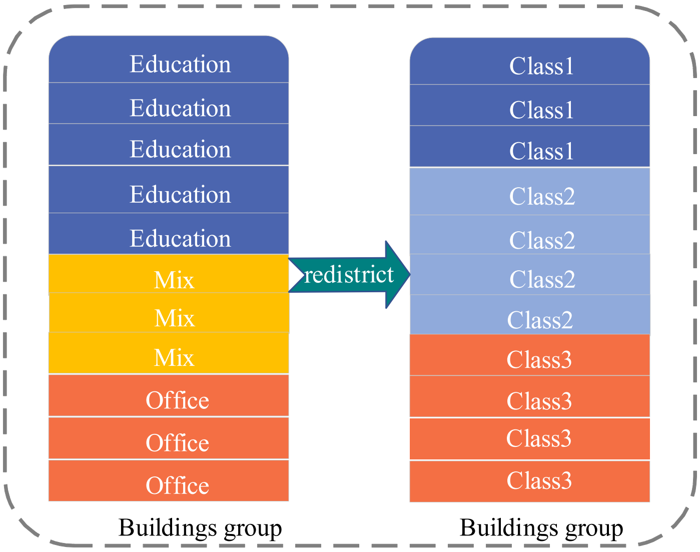

**Figure 9.** Schematic diagram of building reclassification.

## 4. Results and Discussion

After the clustering is complete, we have a new classification of buildings. We compared the original classification with the methods proposed in this paper. To evaluate the performance of the two methods, we calculated the CH and DBI indices of the two grouping methods, respectively. As shown in Table 4, the results show that DBI is smaller and CH is larger in the proposed method. This confirms that the clustering results of the proposed method have higher similarity in intra-group building and greater differences between different clusters than those based on the original classification. Therefore, the energy consumption benchmarking is more in line with the actual situation.

**Table 4.** Cluster evaluation index.

| Index | Original Classification | New Classification |
|---|---|---|
| DBI | 1.647 | 0.645 |
| CH | 81.263 | 355.391 |

Therefore, the benchmarking of building energy consumption obtained solves the problems that the classification of the original campus buildings in the classification model is inconsistent with reality, affected by the number of floors, degree days and other factors, and that mixed-function buildings are not well classified according to the actual situation. The new energy baseline can meet the need for energy conservation benchmarks. The energy consumption quartile lines of three types of buildings are shown in Table 5. In addition, we use the Quartile1, Quartile2 and Quartile3 as the low, medium and high energy benchmarks of building energy consumption.

**Table 5.** Energy consumption quartile lines of the building.

| Building Classification | EUI/(KW·h/m$^2$) | | |
|---|---|---|---|
| | Quartile 1 | Quartile 2 | Quartile 3 |
| Class 1 | 51.93 | 87.02 | 160.24 |
| Class 2 | 115.79 | 145.54 | 186.90 |
| Class 3 | 153.16 | 176.38 | 191.46 |

## 5. Conclusions

In this study, we analyzed the energy consumption data of more than 200 buildings in colleges and universities, aiming to provide some help for the determination method of energy consumption benchmarking. At the same time, the case of this paper also provides information to support the energy consumption benchmarking of colleges and universities. Specifically speaking, this paper includes the following contents:

(1) The random forest model was used to determine several main characteristics affecting building energy consumption and building classification. Considering building EUI and building the original classification label as double objectives, the random forest model was constructed successively to obtain the importance ranking of building features' contribution to building EUI and building original classification, so that the buildings in the obtained clustering result not only adhere to mathematical laws of relation but also have the necessary similarities in practical work. In this way, the building features that have an important influence on both the EUI of buildings and the original classification of buildings can be obtained, which lays a foundation for further making improving a fine energy-saving benchmarking.

(2) The dimensionality of the important building features obtained in the above steps is reduced to the building cluster type to eliminate errors by factor analysis. The K-means method is adopted for cluster analysis of the building set, and the common factors extracted from the campus buildings are clustered to remove the influence of each building feature on the energy consumption level in the classification. We aimed to solve the main problem, which was that the original building classification is not practical. Thus, the energy consumption reference line measured by the quartile method is more practical value. For each kind of building, three levels of low, medium and high energy consumption level are proposed, respectively. It makes the method of evaluating the energy use level and energy saving potential of campus buildings much more reasonable.

We analyzed the data on the energy consumption of campus buildings, provided a reference for the delineation of energy consumption lines of campus buildings in the future and also provided case support for the energy consumption benchmarking of universities. On the basis of the research results obtained in this paper, future research can extend the data-driven building energy consumption assessment technology to more building features and more building classification, and obtain the key influencing factors of energy consumption of different types of buildings through data mining technology at a higher dimension, so as to obtain more accurate energy consumption benchmarking.

**Author Contributions:** Conceptualization, B.W.; methodology, B.W.; software, B.W.; validation, B.W. and H.Z.; formal analysis, B.W.; writing—original draft preparation, B.W. and Z.Z.; writing—review and editing, B.W., T.W. and D.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Acknowledgments:** The author is grateful to the editors and anonymous referees for their comments and suggestions.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Zhu, H.; Goh, H.H.; Zhang, D.; Ahmad, T.; Liu, H.; Wang, S.; Li, S.; Liu, T.; Dai, H.; Wu, T. Key technologies for smart energy systems: Recent developments, challenges, and research opportunities in the context of carbon neutrality. *J. Clean. Prod.* **2021**, *331*, 129809. [CrossRef]
2. Lei, L.; Wu, B.; Fang, X.; Chen, L.; Wu, H.; Liu, W. A dynamic anomaly detection method of building energy consumption based on data mining technology. *Energy* **2023**, *263*, 125575. [CrossRef]
3. Tong, P.; Zhao, C.; Wang, H. Research on the Survival and Sustainable Development of Small and Medium-Sized Enterprises in China under the Background of Low-Carbon Economy. *Sustainability* **2019**, *11*, 1221. [CrossRef]
4. Zhang, D.; Li, H.; Zhu, H.; Zhang, H.; Goh, H.H.; Wong, M.C.; Wu, T. Impact of COVID-19 on Urban Energy Consumption of Commercial Tourism City. *Sustain. Cities Soc.* **2021**, *73*, 103133. [CrossRef]
5. 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]
6. Pham, A.D.; Ngo, N.T.; Truong, T.T.H.; Huynh, N.T.; Truong, N.S. Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *J. Clean. Prod.* **2020**, *260*, 121082. [CrossRef]
7. Xu, Z.; Kuishan, L. Sensitivity analysis of several energy-saving measures to residential building energy consumption in hot summer and cold winter area. *HVAC* **2008**, *38*, 6.
8. Ma, J.; Cheng, J.C.P. Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests. *Appl. Energy* **2016**, *183*, 193–201. [CrossRef]
9. Hsu, D. Identifying key variables and interactions in statistical models of building energy consumption using regularization. *Energy* **2015**, *83*, 144–155. [CrossRef]
10. Robinson, C.; Dilkina, B.; Hubbs, J.; Zhang, W.; Guhathakurta, S.; Brown, M.A.; Pendyala, R.M. Machine learning approaches for estimating commercial building energy consumption. *Appl. Energy* **2017**, *208*, 889–904. [CrossRef]
11. US Energy Star Portfolio Manager. Energy Star Score. Technical Reference. 2019. Available online: https://portfoliomanager.energystar.gov/pdf/reference/ENERGY%20STAR%20Score.pdf (accessed on 3 July 2020).
12. Walker, S.; Khan, W.; Katic, K.; Maassen, W.; Zeiler, W. Accuracy of different machine learning algorithms and added-value of predicting aggre-gated-level energy performance of commercial buildings. *Energy Build.* **2020**, *209*, 109705. [CrossRef]
13. Guo, Y.; Wang, J.; Chen, H.; Li, G.; Liu, J.; Xu, C.; Huang, R.; Huang, Y. Machine learning-based thermal response time ahead energy demand prediction for building heating systems. *Appl. Energy* **2018**, *221*, 16–27. [CrossRef]
14. Ascione, F.; Bianco, N.; De Stasio, C.; Mauro, G.M.; Vanoli, G.P. Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: A novel approach. *Energy* **2017**, *118*, 999–1017. [CrossRef]
15. Bui, D.K.; Nguyen, T.N.; Ngo, T.D.; Nguyen-Xuan, H. An artificial neural network (ANN) expert system enhanced with the electromag-netism-based firefly algorithm (EFA) for predicting the energy consumption in buildings. *Energy* **2020**, *190*, 116370. [CrossRef]
16. Kim, M.; Jung, S.; Kang, J. Artificial neural network-based residential energy consumption prediction models considering resi-dential building information and user features in South Korea. *Sustainability* **2019**, *12*, 109. [CrossRef]

17. Wei, Z.; Zou, Y.; Wang, H. Introduction on British public building energy consumption benchmark evaluation methods and en-lightenment for China. *Build. Sci.* **2011**, *27*, 7–12.
18. Yong, C.; Zheng, W.; Hui, L.; Chong, M. The Inspiration of German VDI 3807 Standard to China's Energy Consumption Quota. *Constr. Sci. Technol.* **2011**, *22*, 78–81.
19. Santamouris, M.; Mihalakakou, G.; Patargias, P.; Gaitani, N.; Sfakianaki, K.; Papaglastra, M.; Pavlou, C.; Doukas, P.; Primikiri, E.; Geros, V.; et al. Using intelligent clustering techniques to classify the energy performance of school buildings. *Energy Build.* **2007**, *39*, 45–51. [CrossRef]
20. Hernandez, P.; Burke, K.; Lewis, J.O. Development of energy performance benchmarks and building energy ratings for non-domestic buildings: An example for Irish primary schools. *Energy Build.* **2007**, *40*, 249–254. [CrossRef]
21. Vaisi, S.; Varmazyari, P.; Esfandiari, M.; Sharbaf, S.A. Developing a multi-level energy benchmarking and certification system for office buildings in a cold climate region. *Appl. Energy* **2023**, *336*, 120824. [CrossRef]
22. Marrone, P.; Gori, P.; Asdrubali, F.; Evangelisti, L.; Calcagnini, L.; Grazieschi, G. Energy benchmarking in educational buildings through cluster analysis of energy retrofitting. *Energies* **2018**, *11*, 649. [CrossRef]
23. *GB/T51161-2016*; Ministry of Housing and Urban-Rural Development, PRC. Energy Consumption Standards for Civil Buildings. China Building and Building Press: Beijing, China, 2016.
24. Zhao, J.; Xin, Y.; Tong, D. Energy consumption quota of public buildings based on statistical analysis. *Energy Policy* **2012**, *43*, 362–370. [CrossRef]
25. Zhu, N.; Zhu, T.L.; Tong, D.D. Determination method of building energy consumption base line of campus degree days method in cold area. *J. Chongqing Univ.* **2016**, *1*, 105–112.
26. Kuo CF, J.; Lin, C.H.; Lee, M.H. Analyze the the energy consumption characteristics and affecting factors of Taiwan's convenience stores-using the big data mining approach. *Energy Build.* **2018**, *168*, 120–136.
27. Deb, C.; Lee, S.E. Determining key variables influencing energy consumption in office buildings through cluster analysis of pre-and post-retrofit building data. *Energy Build.* **2018**, *159*, 228–245. [CrossRef]
28. Li, S.; Chen, Y. Internal benchmarking of higher education buildings using the floor-area percentages of different space usages. *Energy Build.* **2020**, *231*, 110574. [CrossRef]
29. Xie, H.; Jiang, M.; Zhang, D.; Goh, H.H.; Ahmad, T.; Liu, H.; Liu, T.; Wang, S.; Wu, T. IntelliSense technology in the new power systems. *Renew. Sustain. Energy Rev.* **2023**, *177*, 113229. [CrossRef]
30. Liu, S.; Zhou, C.; Guo, H.; Shi, Q.; Song, T.E.; Schomer, I.; Liu, Y. Operational optimization of a building-level integrated energy system considering additional potential benefits of energy storage. *Prot. Control. Mod. Power Syst.* **2021**, *6*, 55–64. [CrossRef]
31. Fan, G.F.; Zhang, L.Z.; Yu, M.; Hong, W.-C.; Dong, S.-Q. Applications of random forest in multivariable response surface for short-term load fore-casting. *Int. J. Electr. Power Energy Syst.* **2022**, *139*, 108073. [CrossRef]
32. Ilbeigi, M.; Ghomeishi, M.; Dehghanbanadaki, A. Prediction and optimization of energy consumption in an office building using artificial neural network and a genetic algorithm. *Sustain. Cities Soc.* **2020**, *61*, 102325. [CrossRef]
33. Kai, Y.; Yan, H.; Kang, L. Significance Scores of Random Forest Variables and Their Research Progress. Available online: https://www.paper.edu.cn (accessed on 12 July 2015).
34. Deng, H.; Fannon, D.; Eckelman, M.J. Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. *Energy Build.* **2018**, *163*, 34–43. [CrossRef]
35. Rodriguez-Galiano, V.; Sanchez-Castillo, M.; Chica-Olmo, M.; Chica-Rivas, M.J.O.G.R. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* **2015**, *71*, 804–818. [CrossRef]
36. Jang, S.; Elmqvist, N.; Ramani, K. MotionFlow: Visual Abstraction and Aggregation of Sequential Patterns in Human Motion Tracking Data. *IEEE Trans. Vis. Comput. Graph.* **2015**, *22*, 21–30. [CrossRef] [PubMed]
37. Hong, T.; Yang, L.; Hill, D.; Feng, W. Data and analytics to inform energy retrofit of high performance buildings. *Appl. Energy* **2014**, *126*, 90–106. [CrossRef]
38. US Energy Information Administration (EIA). Available online: https://www.eia.gov/ (accessed on 3 September 2022).
39. Yang, J.; Ning, C.; Deb, C.; Zhang, F.; Cheong, D.; Lee, S.E.; Sekhar, C.; Tham, K.W. k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy Build.* **2017**, *146*, 27–37. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.