Application of the BORUTA Algorithm to Input Data Selection for a Model Based on Rough Set Theory (RST) to Prediction Energy Consumption for Building Heating

Tomasz Szul 1,*, Sylwester Tabor 1 and Krzysztof Pancerz 2

1 Faculty of Production and Power Engineering, University of Agriculture, 30-149 Krakow, Poland; s.tabor@urk.edu.pl
2 Department of Technology and Computer Science, Szymon Szymonowic State School of Higher Education in Zamość, 22-400 Zamość, Poland; krzysztof.pancerz@upz.edu.pl
* Correspondence: t.szul@urk.edu.pl; Tel.: +48-12-662-46-47

Abstract: Energy prediction used for building heating has attracted particular attention because it is often required in the development of various strategies to improve the energy efficiency of buildings, especially those undergoing thermal improvements. The complexity, dynamics, uncertainty, and nonlinearity of existing building energy systems create a great need for modeling techniques. One of them is machine learning models, which are based on input data consisting of features that describe the objects under study. The data describing actual buildings used to build the model may be characterized by missing values, duplicate or inconsistent features, noise, and outliers. Therefore, an extremely important aspect of the prediction model development effort is the proper selection of features to simplify the prediction of energy consumption for heating. In this connection, the goal was to evaluate the usefulness of a model describing the final energy demand rate for building heating using groups of features describing actual residential buildings undergoing thermal retrofit. The model was created by combining two algorithms: the BORUTA feature selection algorithm, which prepares conditional variables corresponding to features for a prediction model based on rough set theory (RST). The research was conducted on a group of 109 multi-family buildings from the end of the last century (made in large-panel technology), thermomodernized at the beginning of the 21st century. Evaluation metrics such as MAPE, MBE, CV RMSE, and $R^2$, which are adopted as statistical calibration standards by ASHRAE, were used to assess the quality of the developed prediction model. The analysis of the obtained results indicated that the model based on RST, based on the features selected by the BORUTA algorithm, gives a satisfactory prediction quality with a limited number of input variables, and thus allows to predict energy consumption (after thermal improvement) for this type of buildings with high accuracy.

Keywords: data selection; BORUTA algorithm; building load prediction; rough set theory; building energy modeling; thermal improvement of buildings
methods are demanding in terms of the domain knowledge as well as computational resources [1,5]. In contrast to it, the machine-data-driven modeling process allows a more flexible approach to the detail, availability, and accuracy of data describing buildings and their surroundings. Finally, data-driven methods, by creating a direct relationship between design parameters and building energy consumption, help to alleviate the heavy computational burden as well as engineering efforts required in evaluating the energy performance of various types of facilities. In building energy system modeling, machine learning techniques have been widely applied both in load prediction [7] as well as in control and operation processes [8]. In general, machine learning algorithms used for heat load/heat demand prediction can be classified into the following types [1,2]:

- linear regression [9–12],
- tree-based [13–15],
- autoregressive methods [16,17],
- neural network [13,18–24],
- deep learning [5,25,26],
- support vector machine [21,26–30],
- hybrid [31–33],
- k-nearest neighbors [34],
- others [35–40].

The mentioned types of machine learning models are based on input data consisting of features that describe the objects under study. A feature is the term for variable, parameter, or attribute in the machine learning language. For data-driven models, feature selection strongly affects model performance. Well-designed features increase flexibility and robustness (i.e., even if a simple algorithm is used, prediction results can still be achieved) which allows simplifying the energy assessment of buildings [1]. The data used to create the model may be characterized by missing values, duplicate or inconsistent features, noise, and outliers. Therefore, the science of feature selection, also known as variable selection or attribute selection, is essential for effective feature identification for a building load prediction model. The following feature selection techniques can be distinguished for making heat demand forecasts [1,41]:

- An embedded method, where feature selection is embedded in machine learning algorithms that include feature selection as an intrinsic module, such as random forest and multivariate adaptive regression with glued functions (MARS) [13,14,41–44].
- A filter method, which belongs to statistical methods, assigning a score to each feature, ranks features according to the score, and finally either retains the feature or removes it from the set of candidate features based on ranking [41,45–49].
- A wrapper method comparing the prediction performance of data-driven models using subsets of candidate features to evaluate the fitness of each subset of candidate features [21,50,51].
- A hybrid method, which involves using at least two of the above techniques as a combined feature selection method [23,52,53].

Most of the studies focus on the development of machine learning algorithms to achieve the lowest possible prediction error [1,2,18]. Authors presenting prediction results for applied machine learning algorithms often only mention what data they use for modeling, but do not provide more details, such as: data resolution, uncertainty, training/test data partitioning, etc. [1]. It can also be noted that the quality and accuracy of prediction models are mostly tested on theoretical datasets obtained from simulations of different types of buildings [4]. There is a lack of solutions for real buildings characterized by different availability of data describing the objects from thermal and operational point of view [1,4,18,35]. An another extremely important aspect when evaluating energy consumption in real buildings is the fact that many input parameters to the model can be obtained based on engineering calculations (e.g., heat transfer coefficients, thermal bridges, ventilation air flows, building use, and the like), which, in many cases, are based
on unreliable and inaccurate data, which significantly affect the accuracy of the building energy assessment [1,4,35]. For real objects, feature selection is an important aspect in building predictive models free from correlated variables, errors, and unwanted noise so feature selection should be done to select the most relevant ones. Therefore, the aim of this paper is to evaluate the usefulness of a model describing the final energy demand rate for heating a building using groups of features describing real residential buildings undergoing thermomodernization. The selection of the listed features was performed using the BORUTA algorithm [54]. Boruta uses the method of selection of all relevant features, where it captures all features that are relevant to the decision variable under certain conditions. The selection results were decided to be used to build a predictive model using rough set theory (RST) [55]. It should be noted that, as shown in the literature review [1], it has not yet been used in the energy assessment of buildings and therefore represents a novelty in this type of research. Given that an important aspect when building a predictive model is to reduce the number of input variables, it was additionally decided to check whether, when building a predictive model, all features indicated by the BORUTA algorithm as relevant should be selected or whether the number of features can be reduced and how such a reduction would affect the accuracy of energy consumption prediction. To assess the quality of the developed prediction model, evaluation metrics such as [56]: \(\text{MAPE}, \text{MBE}, \text{CV RMSE}\) and \(R^2\), which are accepted as statistical calibration standards by ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) [57], were used.

2. Materials and Methods

The research was conducted on a sample of one hundred and nine multi-family residential buildings erected in the 1960s and 1970s in prefabricated technology (made of large plates). The buildings underwent thermomodernization in 2010–2015. The considered subjects are heated from the municipal district heating network. Thanks to that, information about actual heat consumption for heating was obtained. For the analyzed group of buildings, energy characteristics were performed in accordance with the methodology for determining the energy performance of the building [58], applicable in Poland, according to the method based on the actual energy consumption for heating. Energy consumption was determined for the state before and after performing thermal improvement, using heat consumption data from three heating seasons. To exclude seasonal variations, the actual energy consumption values were adjusted to standard season conditions. On this basis, indices of final energy demand for heating were determined. Additionally, for buildings, which did not have information regarding the ordered thermal power for heating, calculations of the design heat load for heating before thermomodernization were carried out according to ISO 12,831 [59]. The analyzed buildings were characterized by a set of features. Some of them are measured and others are calculated. These features are of typical energetic character, like, for example, power demand for heating the building or seasonal energy demand for heating, and others describe for example structural parameters like surface areas of partitions through which heat losses occur, heated volume, heat transfer coefficients through partitions. These features describe buildings in the state before thermal improvements were made. Utility parameters are also taken into account, such as, for example, the number of people using the building. The data set includes also qualitative features, such as information, which partition was subjected to thermomodernization. Each of the qualitative features could be in two response options “Yes” or “No”. These words were replaced by digits: “1” and “0”. The value “1” was put when a given event occurred (the partition was thermomodernized). Otherwise, when the event did not occur, the feature took the value “0”. The features characterizing the investigated buildings were marked successively \(v_1, v_2, \ldots, v_{23}, v_{24}\), and they are conditional variables (corresponding to features) for the rough model. The decision variable (index of final energy demand for heating after modernization) is labeled with \(d\). The features describing the building are
Table 1. Characteristics of features affecting the reduction of annual energy demand for heating in multifamily residential buildings.

| Attribute                                                                 | Unit      | Min       | Max        | Average   | Coefficient of Variation * |
|---------------------------------------------------------------------------|-----------|-----------|------------|-----------|---------------------------|
| v₁–the heated volume of building (EM)                                     | m³        | 1308.0    | 22935.6    | 6393.1    | 62.9                      |
| v₂–total (net internal area) (IM)                                         | m²        | 384.9     | 5077.2     | 1764.0    | 54.4                      |
| v₃–surface of heated floors (IM)                                          | m²        | 372.7     | 4985.0     | 1568.5    | 53.4                      |
| v₄–surface of roof projection area (net) (EM)                             | m²        | 87.2      | 2153.0     | 467.0     | 55.5                      |
| v₅–total walls surface (net) area (EM)                                    | m²        | 296.3     | 3897.3     | 1096.8    | 49.7                      |
| v₆–surface of floor from interior measurements (floor over basement or floor on the ground) (IM) | m²        | 12.2      | 1075.1     | 395.4     | 58.8                      |
| v₇–total windows area (EM)                                                | m²        | 69.7      | 838.3      | 290.7     | 54.1                      |
| v₈–number of residential flats, premises                                  | pc.       | 6.0       | 142.0      | 32.4      | 62.8                      |
| v₉–number of living persons per building                                   | Nb.       | 14.0      | 425.0      | 73.9      | 67.6                      |
| v₁₀–shape coefficient of buildings (the ratio surface to volume)         | m⁻¹       | 0.21      | 1.92       | 0.46      | 43.7                      |
| v₁₁–calculated thermal transmittance of walls components                  | W·m⁻²·K⁻¹ | 0.43      | 1.89       | 1.09      | 31.9                      |
| v₁₂–calculated thermal transmittance of peak walls (side, narrower building walls) components | W·m⁻²·K⁻¹ | 0.32      | 3.09       | 1.01      | 47.7                      |
| v₁₃–calculated thermal transmittance of roof projections components       | W·m⁻²·K⁻¹ | 0.18      | 3.56       | 1.25      | 85.1                      |
| v₁₄–calculated thermal transmittance of floors components (floor over basement) | W·m⁻²·K⁻¹ | 0.26      | 2.43       | 1.17      | 37.8                      |
| v₁₅–calculated thermal transmittance of floor components on the ground    | W·m⁻²·K⁻¹ | 0.38      | 3.51       | 1.61      | 51.7                      |
| v₁₆–thermal transmittance of windows (commercial data)                    | W·m⁻²·K⁻¹ | 1.2       | 3.5        | 1.88      | 30.7                      |
| v₁₇–information whether the wall to be thermal improved                   | -         | 0         | 1          | 0.98      | 13.7                      |
| v₁₈–information whether the peak wall (side, narrower building wall) to be thermal improved | -         | 0         | 1          | 0.93      | 28.3                      |
| v₁₉–information whether the roof to be thermal improved                   | -         | 0         | 1          | 0.52      | 96.0                      |
| v₂₀–information whether the floor over basement to be thermal improved    | -         | 0         | 1          | 0.32      | 146.1                     |
| v₂₁–information whether the floor on the ground to be thermal improved    | -         | 0         | 1          | 0.06      | 383.5                     |
| v₂₂–information whether replacement windows                               | -         | 0         | 1          | 0.06      | 416.2                     |
| v₂₃–heating consumed power                                                | kW        | 36.5      | 413.5      | 129.8     | 57.0                      |
| v₂₄–index of final energy demand for heating before modernization (calculated from measurements of actual energy consumption for heating) | kWh·m⁻²·year⁻¹ | 83.0      | 566.1      | 253.4     | 43.6                      |
| d–index of final energy demand for heating after modernization            | kWh·m⁻²·year⁻¹ | 51.2      | 389.2      | 143.6     | 46.1                      |

Where: (remark: EM denotes calculation from exterior measurements, IM denotes calculation from interior measurements). * Coefficient of variation—it is defined as the quotient (expressed in %) of the absolute measure of variation (standard deviation) and the average level of the trait value (arithmetic mean).

The set of features describing the energy consumption in the investigated buildings presented in Table 1 is characterized by a large variability. The specificity of the features also indicates a great variety of ways of coding the given features, which occur both in qualitative and quantitative form, some of them are measured and some of them are calculated, so reference point is needed to help distinguishing the really important features from the unimportant ones. To cope with this problem, we used an algorithm that provides...
criteria for selecting relevant (important) features. To select relevant features, we have applied the BORUTA algorithm [54]. It is worth noting that in addition to selecting relevant features, the algorithm also creates a ranking of their relevance. The BORUTA algorithm applies random forests for feature relevance estimation. The random forest is a very popular learning model for solving a variety of classification and regression problems [60]. The random forest is based on a multitude of decision trees which are independently developed on different sample bags taken from the training set. The algorithm is designed as a wrapper method. The original dataset is extended by adding the so-called shadow features whose values are randomly permuted among the training cases to remove their correlations with a decision variable. The importance estimation of a feature is calculated as the loss of classification accuracy caused by a random permutation of feature values of cases. First, the loss of classification accuracy is computed individually for all decision trees in the forest which make use of a given feature to classify cases and then the average and standard deviation of the loss of classification accuracy are computed. An importance measure is the Z score computed by dividing the average loss by its standard deviation. The importance measure is used to determine the ranking of features. In addition to generation of the feature ranking, the BORUTA algorithm classifies features into three types:

- Confirmed.
- Tentative.
- Rejected.

The Z score is calculated for each feature. Then, the maximum Z score (MZS) among shadow features is identified and a hit is assigned to every feature that is scored better than MZS. The two-sided equality test with MZS is applied. The features which have importance significantly lower than MZS are treated as irrelevant (rejected) ones. The features which have importance significantly higher than MZS are treated as relevant (confirmed) ones. The remaining features are treated as tentative ones.

After establishing the list of “importance” of features, the developed database was subdivided into a training set, to which 80% of the studied buildings were randomly selected, and a test set formed from the remaining objects. A method based on rough set theory (RST) was used to build a model to determine the annual heat demand of residential buildings undergoing thermal upgrading. Approximate sets proposed by Z. Pawlak [56] are a suitable tool for dealing with general (imprecise) and ambiguous data. The rough sets serve as a methodology in the process of discovering knowledge in data-bases. It is a tool used to describe inaccurate, uncertain knowledge; to model decision-making systems; and for approximation reasoning [55]. It allows us to bring more flexibility in data mining to rough set theory and to analyze observations expressed in quantitative and qualitative forms. The methodology for building predictive models based on rough set theory is presented in papers [35,61–63]. To build the model, the rough set exploration system RSES 2.1 (Department of Logic of the Institute of Mathematics, University of Warsaw, Poland) was used, which is a computer tool that enables the analysis of data in a tabular form, using rough set theory [64].

To evaluate the quality of the developed prediction model, evaluation metrics such as [56] mean bias error (MBE), coefficient of variance root mean square error (CV RMSE) and coefficient of determination ($R^2$), and mean absolute percentage error (MAPE), were used.

\[ MBE = \frac{\sum_{m=1}^{n_g} (y_i - y_i')}{\sum_{i=1}^{n_g} y_i}.100\% \quad m = 1, 2, 3 \ldots, n_g \]  
\[ CV \ RMSE = \sqrt{\frac{\sum_{m=1}^{n_g} (y_i - y_i')^2}{\sum_{i=1}^{n_g} y_i^2}}.100\% \quad m = 1, 2, 3 \ldots, n_g \]
where \( y_i \) is the actual value (quantity) in the facility \( i \), and \( y'_i \) is the forecast value (quantity) in the facility \( i \). The difference between \( y_i \) and \( y'_i \) is divided by the actual value \( y_i \) and \( m \) is the index of the test object, \( n_g \) is the number of objects (\( m = 1,2,3,...,n_g \)).

According to ASHRAE Guideline 14 criteria [56,57], for the model to be considered well calibrated, the value of the evaluation indices should not exceed:

- MBE index \( \pm 5\% \),
- CV RMSE index 15%.

However, the value of the coefficient of determination should be \( R^2 \geq 0.75 \).

To see how the choice of features would affect the prediction accuracy of the model, it was decided to create four subsets of input features for comparison, i.e.,

- SET 1—the subset of features indicated by the algorithm as “confirmed”.
- SET 2—a subset of the features indicated as “confirmed” by the algorithm, with the top 10 (indicated by the algorithm) features selected for analysis.
- SET 3—a subset of the features indicated as “confirmed” by the algorithm, with the top 3 (indicated by the algorithm) features selected for analysis.
- SET 4—the subset of features indicated by the algorithm as “confirmed” + “tentative”.

### 3. Results and Discussion

In the first part of building the prediction model, the database containing 24 features was subjected to selection to find all of the important features influencing the final energy consumption after thermomodernization. The listed features inform about e.g., power demand, seasonal energy consumption for heating of the building before thermomodernization, area of partitions, through which heat losses occur, heated cubature, heat transfer coefficients, number of occupants as well as which partition was thermomodernized. The selection of the mentioned features was performed in R using the BORUTA package. Boruta uses all relevant features selection method, where it captures all features that are relevant to the decision variable under certain circumstances. This algorithm finds all features that are strongly or weakly significant for the decision variable. The selection results are summarized in Table 2 and illustrated in box plot 2.

In Table 2, the meaning of columns is as follows: Mean Imp—the mean of \( IM_p \), Median Imp—the median of \( IM_p \), Min Imp—the minimum of \( IM_p \), Max Imp—the maximum of \( IM_p \), Norm Hits - the number of hits normalized to number of importance source runs, where \( IM_p \) is the importance measure computed over multiple iterations. As it can be seen in Table 2, BORUTA indicated the answer on the importance of features in the dataset. In this case, out of 24 features, six of them are rejected and 15 are confirmed. Three features are marked as tentative. The tentative features have meanings so close to their best shadow features, the Boruta is unable to make decisions with the desired confidence in the default number of runs of the random forest.

The resulting graph (Figure 1) generated using the BORUTA package in the R environment, after features classification, indicated the importance (axis Y) of the analyzed features (put on axis X) by ranking and color-coding them. From the box-plot, the blue box-plots correspond to the shadow features. We have three blue box-plots for minimum, average, and maximum shadow feature values. The green box-plots correspond to features that were confirmed as valid features, and the red box-plots correspond to features that
were confirmed as irrelevant features. The yellow box-plots correspond to features that are uncertain, i.e., the algorithm was not able to come to a conclusion about their importance.

Table 2. The result of the feature selection process.

| No. | Attribute | Mean Imp | Median Imp | Min Imp | Max Imp | NormHits | Decision |
|-----|-----------|----------|------------|---------|---------|----------|----------|
| 1   | v24       | 17.906   | 18.077     | 15.326  | 19.881  | 1        | Confirmed |
| 2   | v10       | 7.854    | 7.816      | 5.915   | 9.887   | 1        | Confirmed |
| 3   | v12       | 7.672    | 7.64       | 6.404   | 9.203   | 1        | Confirmed |
| 4   | v2        | 6.461    | 6.466      | 5.071   | 7.989   | 0.97     | Confirmed |
| 5   | v3        | 6.433    | 6.455      | 4.668   | 7.938   | 0.97     | Confirmed |
| 6   | v15       | 6.278    | 6.32       | 4.626   | 8.065   | 0.96     | Confirmed |
| 7   | v13       | 5.26     | 5.225      | 2.953   | 7.372   | 0.95     | Confirmed |
| 8   | v11       | 4.934    | 4.999      | 2.242   | 6.756   | 0.92     | Confirmed |
| 9   | v4        | 4.893    | 4.929      | 3.23    | 6.224   | 0.9    | Confirmed |
| 10  | v15       | 6.278    | 6.32       | 4.626   | 8.065   | 0.96     | Confirmed |
| 11  | v18       | 4.359    | 4.373      | 2.278   | 5.549   | 0.879    | Confirmed |
| 12  | v1        | 4.279    | 4.374      | 2.259   | 5.923   | 0.899    | Confirmed |
| 13  | v6        | 3.958    | 3.952      | 1.88    | 5.736   | 0.859    | Confirmed |
| 14  | v7        | 3.632    | 3.724      | 1.079   | 5.457   | 0.778    | Confirmed |
| 15  | v14       | 3.234    | 3.242      | 0.814   | 4.884   | 0.687    | Confirmed |
| 16  | v19       | 2.962    | 2.951      | 1.332   | 4.122   | 0.637    | Confirmed |
| 17  | v16       | 2.805    | 2.821      | 0.447   | 4.563   | 0.607    | Tentative |
| 18  | v5        | 2.678    | 2.78       | −0.413  | 4.376   | 0.607    | Tentative |
| 19  | v9        | 1.861    | 1.805      | −0.711  | 4.501   | 0.304    | Rejected  |
| 20  | v8        | 1.658    | 1.595      | 0.609   | 3.157   | 0.031    | Rejected  |
| 21  | v20       | 1.602    | 1.555      | −0.732  | 3.316   | 0.213    | Rejected  |
| 22  | v21       | 0.939    | 1.252      | −0.669  | 2.113   | 0.021    | Rejected  |
| 23  | v22       | −0.48    | −0.608     | −1.964  | 0.777   | 0        | Rejected  |
| 24  | v17       | −0.684   | −1.001     | −1.926  | 1.002   | 0        | Rejected  |

Figure 1. Feature selection importance graph using the BORUTA algorithm.
Based on the selection result, it can be concluded that the important features (confirmed), that affect the final energy consumption after thermal upgrading, are: v24–index of final energy demand for heating before modernization (calculated from measurements of actual energy consumption for heating), v10–shape coefficient of buildings (the ratio surface to volume), v12–calculated thermal transmittance of peak walls components, v2–calculated from interior measurements total (net internal area), v3–calculated surface of heated floors from interior measurements, v15–calculated thermal transmittance of floor components on the ground, v13–calculated thermal transmittance of roof projections components, v11–calculated thermal transmittance of walls components, v4–calculated from exterior measurements surface of roof projection area (net), v23–heating consumed power, v18–information whether the peak wall to be thermal improved, v1–calculated from exterior measurements the heated volume of building, v6–calculated surface of floor from interior measurements (floor over basement or floor on the ground), v7–calculated from exterior measurements total windows area, v14–calculated thermal transmittance of floors components (floor over basement).

However, it is uncertain (but cannot be excluded) whether features such as v19, information whether the roof needs to be thermal improved, v16, thermal transmittance of windows, and v5, calculated from exterior measurements total wall surface area, have an impact on energy consumption after thermal upgrading.

After the selection of features, the next step in building the predictive model was to randomly divide the database into a training set and a test set (in an 80/20 ratio). Considering the obtained feature selection results, the input data characterizing the buildings were divided into four sets of features, which differ in the number of features, i.e., set 1, which contained the data indicated by the algorithm as “confirmed”, the top 10 features indicated by the algorithm as “confirmed” for which the Norm Hits index value was greater than 0.9 were selected for set 2. In the third set, we limited ourselves to 3 features for which the Norm Hits value was equal to 1 and set 4, which consists of the features labeled as “confirmed” and “tentative”. Table 3 lists the individual features included in the sets of features, with the feature’s membership in a given set denoted by 1.

| No. | Feature | Sets of Features |
|-----|---------|------------------|
|     |         | Set 1 | Set 2 | Set 3 | Set 4 |
| 1   | v24     | 1     | 1     | 1     | 1     |
| 2   | v10     | 1     | 1     | 1     | 1     |
| 3   | v12     | 1     | 1     | 1     | 1     |
| 4   | v2      | 1     | 1     | 1     | 1     |
| 5   | v3      | 1     | 1     | 1     | 1     |
| 6   | v15     | 1     | 1     | 1     | 1     |
| 7   | v13     | 1     | 1     | 1     | 1     |
| 8   | v11     | 1     | 1     | 1     | 1     |
| 9   | v4      | 1     | 1     | 1     | 1     |
| 10  | v23     | 1     | 1     | 1     | 1     |
| 11  | v18     | 1     | 1     | 1     | 1     |
| 12  | v1      | 1     | 1     | 1     | 1     |
| 13  | v6      | 1     | 1     | 1     | 1     |
| 14  | v7      | 1     | 1     | 1     | 1     |
| 15  | v14     | 1     | 1     | 1     | 1     |
| 16  | v19     | 1     | 1     | 1     | 1     |
| 17  | v16     | 1     | 1     | 1     | 1     |
| 18  | v5      | 1     | 1     | 1     | 1     |

The sets of features presented in Table 3 will allow one to see whether in building a predictive model one should select all features identified as “confirmed” by the BORUTA algorithm, or whether one can limit the number of features to the group for which the
Norm Hits index value will be greater than 0.9 or equal to 1, and how such a limitation will affect the accuracy of energy consumption prediction. Evaluating the features summarized in Set 4 will further answer the question of how the features marked as tentative will answer affect the prediction accuracy of the model. The features selected by the BORUTA algorithm, describing buildings in the state before thermomodernization, were used to build a model for predicting energy consumption for heating after thermomodernization, which is based on rough set theory (RST). For each group from the training set, reducts and the core of the feature set were determined and used to build the inference model. The reduct in rough set theory is a sufficient set of features that can be used to reason about the value of the decision variable. For a given data set, many reducts can be found. Therefore, the set of features which is common to all reducts can be determined. This set is called the core. The built model was critically analyzed (on the test set) for the accuracy and quality of the built prediction using evaluation metrics such as MAPE, MBE, CV RMSE and $R^2$ [57]. The results of calculating the quality and accuracy of the built models depending on the selected set of features are presented in Table 4.

Table 4. Evaluation of the prediction model of heating energy demand ratio based on the tested sets of features emerged by BORUTA algorithm.

| Assessment Parameters | Set 1 | Set 2 | Set 3 | Set 4 |
|-----------------------|-------|-------|-------|-------|
| MAPE (%)              | 9.2   | 10.7  | 10.3  | 14.2  |
| MBE (%)               | 2.4   | 3.6   | 4.7   | 0.48  |
| CV RMSE (%)           | 5.49  | 5.51  | 5.89  | 5.92  |
| $R^2$ (-)             | 0.85  | 0.82  | 0.81  | 0.79  |

Analyzing the results presented in Table 4, it can be concluded that the model based on rough set theory (RST), based on the features selected by the BORUTA algorithm, allows to obtain satisfactory prediction quality. However, as it can be seen, dataset 1 (marked as “confirmed” by the BORUTA algorithm) affects the better fit of the model to the real data-this is confirmed by the values of three evaluation indicators. Limiting the number of features to ten for which Norm Hits > 0.9 (dataset 2) had little effect on changing the magnitude of the evaluation indices (MBE and MAPE error values increased by about 1%, CV RSME error increased by 0.02%, while $R^2$ coefficient of determination decreased by 0.03). Similarly, results were observed when the selection of three features indicated by the algorithm as the best were selected: Norm Hits = 1 (Set 3), MAPE error values increased by about 1%, MBE by 2.3%, CV RSME error increased by 0.4%, while $R^2$ coefficient of determination decreased by 0.04 compared to the first set of features. Thus, the model based on a limited number of features (Set 2 and 3) can be considered suitable for practical application. Set 4, which in the set of features also includes data marked as “uncertain” by BORUTA, gives worse predictive results (only one MBE index is more favorable than in Set 1). This indicates that the increased number of features negatively affected the predictive quality of the model. The presented prediction results based on features selected by the algorithm give significantly better results compared to the model built based on manual feature selection based on domain knowledge. In papers [35], based on the same set of real objects, a prediction models was built based on 4 sets of features with different amounts of data that were manually selected. In this case, the best result was characterized by error values of for a model based on rough set theory (RST): MAPE 14%, MBE –9.6%, and CV RMSE 18.8%

4. Conclusions and Perspectives

Based on analyses carried out on a group of 109 multi-family residential buildings constructed in large-panel technology, for which energy characteristics were prepared based on the actual energy consumption for heating before and after performing thermal
improvement, a group of 24 features was distinguished. These data can be used to build a prediction model of energy consumption for heating after thermal improvement. Since features describing real buildings may be characterized by missing values, duplicate or inconsistent features, noise, and outliers, it was decided to perform a selection of features to choose the most relevant ones. The selection of the mentioned features was performed using the BORUTA algorithm. Boruta uses all significant features selection method, where he captures all features that are significant for the decision variable under certain conditions.

It was decided to use the selection results to build a prediction model using rough set theory (RST). Additionally, it was decided to check whether building a prediction model one should select all the features defined by the BORUTA algorithm as “confirmed” or whether one can limit the number of features, and how such a limitation would influence the accuracy of energy consumption prediction. The obtained prediction model quality results were evaluated according to statistical calibration standards that are adopted by ASHRAE. Based on the calculations performed, the following conclusions can be made:

- Evaluating the usefulness of the different sets of variables according to the indices proposed by ASHARE, it was found that the use of the feature sets indicated by the BORUTA algorithm in the prediction model yielded errors for the test set of: MAPE 9.2–10.7%; MBE 2.4–4.7%; CV RMSE 5.49–5.89%; and $R^2$ 0.85–0.81,
- The data set consisting of 14 descriptive characteristics marked by the BORUTA algorithm as “confirmed” is the best fit of the model to the real data, confirmed by the values of all evaluation indicators,
- Limiting the number of conditional attributes to 10, had little effect on changing the size of the evaluation indices. Analogous results were obtained when selecting 3 conditional attributes indicated by the algorithm as the best, so the model based on a limited number of input variables can be considered suitable for practical application,
- The method of data classification indicated in this paper together with the forecasting model will make it possible to quickly determine the energy saving potential of buildings without the need to perform detailed (and thus expensive) engineering calculations. Three data describing the building before thermomodernization can be sufficient to estimate energy consumption (after thermal improvement): index of final energy demand for heating before modernization (calculated from measurements of actual energy consumption for heating), shape coefficient of buildings (the ratio surface to volume) and calculated thermal transmittance of peak walls components,
- Increasing the number of conditional attributes beyond those indicated as “confirmed” by the BORUTA algorithm yields poorer predictive performance. This indicates that an increased number of attributes negatively affects the predictive quality of the model,
- In further research, the authors plan to test the usefulness of the presented set of methods for predicting energy consumption in other types of real buildings, such as single-family residential buildings, schools, kindergartens, and others. This will allow a broader assessment of the usefulness of the BORUTA algorithm for input data selection in an energy consumption forecasting model based on rough set theory. It is also planned to compare other methods of classification of input variables based on these objects, i.e., real ones,
- To be able to compare the quality of the predictions with the results in other works, the authors plan to use publicly available databases such as “ASHRAE - Great Energy Predictor III, the Kaggle competition” [65] to evaluate the performance of the algorithms. This will allow to test and compare the applicability of the presented method on data that are used by other researchers in building prediction models.

**Author Contributions:** Conceptualization, T.S.; software, K.P. and T.S.; data curation, T.S.; investigation, T.S.; methodology, T.S.; project administration, T.S.; supervision, T.S.; funding, S.T.; writing—original draft, T.S.; writing—reviewing and editing T.S. and S.T. All authors have read and agreed to the published version of the manuscript.
Funding: This research was funded by University of Agriculture in Krakow, Poland.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article: doi:10.3390/en13061309 (accessed on 2 March 2021).

Acknowledgments: We are grateful to Thomas G. Mathia from Laboratoire de Tribologie et Dynamique des Systèmes, École Centrale de Lyon, France for very fruitful scientific discussions.

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

References

1. Zhang, L.; Wen, J.; Li, Y.; Chen, J.; Ye, Y.; Fu, Y.; Livingood, W. A review of machine learning in building load prediction. Appl. Energy 2021, 285, 116452. [CrossRef]
2. Li, X.; Wen, J. Review of building energy modeling for control and operation. Renew. Sustain. Energy Rev. 2014, 37, 517–537. [CrossRef]
3. Zhao, H.; Magoulès, F. A review on the prediction of building energy consumption. Renew. Sustain. Energy Rev. 2012, 16, 3586–3592. [CrossRef]
4. Bourdeau, M.; Zhai, X.-Q.; Nefzaoui, E.; Guo, X.; Chatellier, P. Modelling and forecasting building energy consumption: A review of data-driven techniques. Sustain. Sustain. Cities Soc. 2019, 48, 101533. [CrossRef]
5. Fumo, N. A review on the basics of building energy estimation. Renew. Sustain. Energy Rev. 2014, 31, 53–60. [CrossRef]
6. Foucquier, A.; Robert, S.; Saund, F.; Stephan, L.; Jay, A. State of the art in building modelling and energy performances prediction: A review. Renew. Sustain. Energy Rev. 2013, 23, 272–288. [CrossRef]
7. Fan, C.; Xiao, F.; Zhao, Y. A short-term building cooling load prediction method using deep learning algorithms. Appl. Energy 2017, 195, 222–233. [CrossRef]
8. Peng, Y.; Rysane, A.; Nagy, Z.; Schlüter, A. Using machine learning techniques for occupancy-prediction-based cooling control in office buildings. Appl. Energy 2018, 211, 1343–1358. [CrossRef]
9. Chou, J.-S.; Bui, D.-K. Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. Energy Build. 2014, 82, 437–446. [CrossRef]
10. Aydinalp-Koksal, M.; Ugursal, V.I. Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector. Appl. Energy 2008, 85, 271–296. [CrossRef]
11. Hygh, J.S.; DeCarolis, J.F.; Hill, D.B.; Ranjithan, S.R. Multivariate regression as an energy assessment tool in early building design. Build. Environ. 2012, 57, 165–175. [CrossRef]
12. Azadeh, A.; Saberi, M.; Seraj, O. An integrated fuzzy regression algorithm for energy consumption estimation with non-stationary data: A case study of Iran. Energy 2010, 35, 2351–2366. [CrossRef]
13. Ahmad, M.W.; Moursheed, M.; Rezgui, Y. Trees vs. Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy Build. 2017, 147, 77–89. [CrossRef]
14. Yu, Z.; Haghighat, F.; Fung, B.C.M.; Yoshino, H. A decision tree method for building energy demand modeling. Energy Build. 2010, 42, 1637–1646. [CrossRef]
15. Ahmad, T.; Chen, H. Short and medium-term forecasting of cooling and heating load demand in building environment with data-mining based approaches. Energy Build. 2018, 166, 460–476. [CrossRef]
16. Yun, K.; Luck, R.; Mago, P.J.; Cho, H. Building hourly thermal load prediction using an indexed ARX model. Energy Build. 2012, 54, 225–233. [CrossRef]
17. Yao, Y.; Lian, Z.; Liu, S.; Hou, Z. Hourly cooling load prediction by a combined forecasting model based on Analytic Hierarchy Process. Int. J. Therm. Sci. 2004, 43, 1107–1118. [CrossRef]
18. Szul, T.; Necka, K.; Mathia, T.G. Neural Methods Comparison for Prediction of Heating Energy Based on Few Hundreds Enhanced Buildings in Four Season’s Climate. Energies 2020, 13, 5453. [CrossRef]
19. Ekici, B.B.; Aksoy, U.T. Prediction of building energy needs in early stage of design by using ANFIS. Expert Syst. Appl. 2011, 38, 5352–5358. [CrossRef]
20. Sholahudin, S.; Han, H. Simplified dynamic neural network model to predict heating load of a building using Taguchi method. Energy 2016, 115, 1672–1678. [CrossRef]
21. Massana, J.; Pous, C.; Burgas, L.; Melendez, J.; Colomer, J. Short-term load forecasting in a non-residential building contrasting models and attributes. Energy Build. 2015, 92, 322–330. [CrossRef]
22. Li, K.; Su, H. Forecasting building energy consumption with hybrid genetic algorithm-hierarchical adaptive network-based fuzzy inference system. Energy Build. 2010, 42, 2070–2076. [CrossRef]
23. Mohammadi, M.; Talebpour, F.; Safaee, E.; Ghadimi, N.; Abedinia, O. Small-Scale Building Load Forecast based on Hybrid Forecast Engine. Neural Process. Lett. 2018, 48, 329–351. [CrossRef]
53. 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]

54. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* 2010, 36, 11. [CrossRef]

55. Pawlak, Z. *Rough Sets. Theoretical Aspects of Reasoning about Data*; Kluwer Academic Press: Dordrecht, The Netherlands, 2012; Available online: http://bcpw.bg.pw.edu.pl/Content/2026/RoughSetsRep29.pdf (accessed on 5 November 2020).

56. Ruiz, G.R.; Bandera, C.R. Validation of Calibrated Energy Models: Common Errors. *Energies* 2017, 10, 1587. [CrossRef]

57. ASHRAE. American Society of Heating, Ventilating, and Air Conditioning Engineers (ASHRAE). *Guideline 14-2014, Measurement of Energy and Demand Savings*; Technical Report; American Society of Heating, Ventilating, and Air Conditioning Engineers: Atlanta, GA, USA, 2014; Available online: https://scholar.google.com/scholar_lookup?title=American+Society+of+Heating,+Ventilating,+and+Air+ Conditioning+Engineers+(ASHRAE).+Guideline+14-2014,+Measurement+of+Energy+and+Demand+Savings&author=ASHRAE&publication_year=2014 (accessed on 22 February 2021).

58. Regulation of the Minister of Infrastructure and Development of 27 February 2015 on the Methodology for Determining the Energy Performance of a Building or Part of a Building and Energy Performance Certificates. Dz.U. 2015 poz. 376. Poland. Available online: http://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU20150000376/O/D20150376.pdf (accessed on 11 February 2021).

59. CEN. *European Standard: Heating Systems in Buildings*; ISO 12831-1:2017-08; CEN: Brussels, Belgium, 2017.

60. Breiman, L. Random forests. *Mach. Learn.* 2001, 45, 5–32. [CrossRef]

61. Nguyen, D.V.; Yamada, K.; Unehara, M. Extended Tolerance Relation to Define a New Rough Set Model in Incomplete Information Systems. *AFS 2013, 372091*. [CrossRef]

62. Renigier-Biłozor, M. Zastosowanie Teorii Zbiorów Przybliżonych do Masowej Wyceny Nieruchomości na Małych Rynkach (Application of Rough Set Theory for Mass Valuation of Real Estate in Small Markets). *Acta Sci. Pol. Adm. Locorum* 2008, 7, 35–51. Available online: http://tnn.org.pl/tnn/publik/19/TNN_Tom_XIX_1.pdf (accessed on 1 March 2021).

63. Szul, T.; Nęcka, K.; Knaga, J. Application of Rough Set Theory to Establish the Amount of Waste in Households in Rural Areas. *Ecol. Chem. Eng. S.* 2017, 24, 311–325. [CrossRef]

64. Rough Set Exploration System 2.1. Available online: http://logic.mimuw.edu.pl/\$\sim\$rses (accessed on 7 January 2021).

65. Miller, C.; Arjunan, P.; Kathirgamanathan, A.; Fu, C.; Roth, J.; Park, J.Y.; Balbach, C.; Gowri, K.; Nagy, Z.; Fontanini, A.D.; et al. The ASHRAE Great Energy Predictor III competition: Overview and results. *Sci. Technol. Built Environ.* 2020, 26, 1427–1447. [CrossRef]