energies

MDPI

Article

Towards an Automated, Fast and Interpretable Estimation Model of Heating Energy Demand: A Data-Driven Approach Exploiting Building Energy Certificates

Antonio Attanasio [1,†], Marco Savino Piscitelli [2,†], Silvia Chiusano [3,*,†], Alfonso Capozzoli [2,†] and Tania Cerquitelli [1,†]

1 Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy; antonio.attanasio@polito.it (A.A.); tania.cerquitelli@polito.it (T.C.)
2 Department of Energy, Politecnico di Torino, 10129 Turin, Italy; marco.piscitelli@polito.it (M.S.P.); alfonso.capozzoli@polito.it (A.C.)
3 Interuniversity Department of Regional and Urban Studies and Planning, Politecnico di Torino, 10129 Turin, Italy
* Correspondence: silvia.chiusano@polito.it; Tel.: +39-011-090-7176
† These authors contributed equally to this work.

Received: 22 February 2019; Accepted: 27 March 2019; Published: 2 April 2019

Abstract: Energy performance certification is an important tool for the assessment and improvement of energy efficiency in buildings. In this context, estimating building energy demand also in a quick and reliable way, for different combinations of building features, is a key issue for architects and engineers who wish, for example, to benchmark the performance of a stock of buildings or optimise a refurbishment strategy. This paper proposes a methodology for (i) the *automatic estimation* of the building *Primary Energy Demand* for space heating (PED_h) and (ii) the *characterization of the relationship* between the PED_h value and the main building features reported by Energy Performance Certificates (EPCs). The proposed methodology relies on a two-layer approach and was developed on a database of almost 90,000 EPCs of flats in the Piedmont region of Italy. First, the *classification layer* estimates the segment of energy demand for a flat. Then, the *regression layer* estimates the PED_h value for the same flat. A different regression model is built for each segment of energy demand. Four different machine learning algorithms (Decision Tree, Support Vector Machine, Random Forest, Artificial Neural Network) are used and compared in both layers. Compared to the current state-of-the-art, this paper brings a contribution in the use of data mining techniques for the asset rating of building performance, introducing a novel approach based on the use of independent data-driven models. Such configuration makes the methodology flexible and adaptable to different EPCs datasets. Experimental results demonstrate that the proposed methodology can estimate the energy demand with reasonable errors, using a small set of building features. Moreover, the use of Decision Tree algorithm enables a concise interpretation of the quantitative rules used for the estimation of the energy demand. The methodology can be useful during both designing and refurbishment of buildings, to quickly estimate the expected building energy demand and set credible targets for improving performance.

Keywords: energy performance certificate; heating energy demand; buildings; data mining; classification; regression; decision tree; support vector machine; random forest; artificial neural network

1. Introduction

Energy efficiency is a growing policy priority for many countries around the world, for both economic and environmental reasons. In the 28 countries that are part of the International Energy Agency (IEA), buildings are responsible for about the 21% of total final energy consumption (26% in Italy) [1]. The amount of this energy used for heating and cooling systems is about 55% in the residential sector (74% in Italy) [1]. Regulatory bodies in several countries took actions to reduce wasteful energy consumption and greenhouse gas emissions and to encourage the use of renewable sources and the design of energy efficient buildings [2].

In most cases, the building energy performance rating has been indicated as a cornerstone to pursue the aforementioned aims. For instance, the *Energy Performance of Buildings Directive* (EPBD), issued by the European Commission, makes the evaluation of energy performance compulsory for new and existing buildings [2].

The EPBD provides member states with guidelines for the building *energy performance certification* process, which includes *energy performance rating* and *energy labeling*. The former is based on a scale of values referred to one or more significant parameters like Energy Use Intensity (EUI) and Primary Energy Demand (PED), while the latter consists in the assignment of an energy performance class (or label) to the building, based on the energy performance rating value. The EPBD lets member states to define the actual implementation of its directives. In Italy the EPBD is currently implemented by various national legislative decrees and technical standards, but there are different rating schemes developed in local areas (regions and autonomous provinces) [3].

Among the existing rating systems worldwide, the *Building Research Establishment's Environmental Assessment Method* (BREEAM) developed in the United Kingdom in 1990, is the first and leading assessment method. *Leadership in Energy and Environmental Design* (LEED) developed in the United States in 1998, is nearly the dominant building assessment system (implemented in more than 40 countries). Other well-known methods include *Comprehensive Assessment System for Building Environmental Efficiency* (CASBEE) of Japan, *National Australian Built Environment Rating System* (NABERS), *Building Environmental Assessment Method of Hong Kong* (HK-BEAM), *Green Mark* of Singapore, *EcoProfile* of Norway, *Deutche Gesellschaft fur Nachhaltiges Bauen* (DGNB) of Germany, *Green Building Label* (GBL) of China [4–6].

The interest in buildings energy performance assessment is increased in the last years, especially to estimate how different features affect the building efficiency. Indeed, from a design perspective, it is very important to determine the effect of the building features on its future energy performance in the early designing phase [7]. Similarly, for existing buildings, it could be useful to evaluate the suitability of a refurbishment plan [8,9]. Whatever the used approach, estimating building energy performance in a quick and reliable way, for different combinations of building features, is a key issue for different actors including public authorities [10]. In this context, Energy Performance Certificate (EPC) provides theoretical measure of how efficient a building could be if operated in standard conditions. However, the performance gap, i.e., the difference between estimated and actual energy performance could be significant. For instance, in [11] is stated that for the Swedish EPCs dataset the assessed performance gap is about the 20% for energy consumption assessments. An EPC is therefore not fully representative of the actual performance during operation but makes it possible to perform comparisons and benchmarking analysis between buildings.

In this paper we propose the *Heating Energy Demand Estimation for Building Asset Rating* (HEDEBAR) methodology providing the following features. (i) HEDEBAR allows the *automatic estimation* of the *Primary Energy Demand for space heating* (PED_h) reported by Energy Performance Certificates (EPCs) (calculated in "standard rating" conditions, according to EN ISO 13790 [12], UNI TS 11300-1 [13], and UNI TS 11300-2 [14]). (ii) Moreover, HEDEBAR allows to unfold the criteria adopted during the asset rating of real buildings, through the extraction of the *principal building features* that contribute *to estimate the building energy demand*. The purpose is twofold: (i) *predictive*, as we define models for the robust energy rating of residential buildings, through the estimation of their PED_h;

(ii) *descriptive*, as we provide an interpretation of the method used to issue EPCs, by highlighting the main features that determine the energy demand of buildings.

The HEDEBAR methodology uses data from EPCs to learn the criteria used by the rating system to issue them. It is based on the hypothesis that building features affect the energy demand in different ways for different classes of building energy efficiency. Therefore, a *two-layer approach* is defined to differentiate the analysis of buildings that belong to distinct *segments of energy demand* (i.e., distinct ranges of PED_h value) and to eventually increase the precision in predicting the PED_h value. In the first layer a *classification* problem is considered to estimate the segment of energy demand of the building to be analyzed. Then, in the second layer a *regression* problem is considered to estimate the PED_h value for the same building. We build a different regression model for each segment of energy demand. The proposed two-layer approach allows us to increase the prediction accuracy with respect to a single layer model, which disregards the possible segment of energy demand of the building.

As a case study, the HEDEBAR methodology has been validated on a dataset of real EPCs of almost 90,000 flats in the Piedmont region of Italy [15–17] released as open data by the Piedmont region. These data are available on a Web platform developed by *CSI Piemonte* (the Information System Consortium) and are regulated by the *Piedmont Region* authority (Sustainable Energy Development Sector).Experimental results obtained on such open data demonstrated that HEDEBAR allows estimating PED_h with a reasonable error by only analyzing a small set of 10 building features. Extracted knowledge, human-readable, can be easily exploited by different stakeholders during the decision making process, e.g., public authorities and regulatory bodies should plan future energy policies that leverage on specific building features [18].

The proposed methodology can be useful for designers and building stakeholders to estimate PED_h and to set reference threshold values for physical input variables. Due to the large dimension of the adopted dataset, the information provided can be considered representative of residential dwelling stock in Piedmont. Moreover, the proposed models are based on statistical variables easy to be adaptable to different datasets. Moreover the developed models can be profitably used by local authorities for a preliminary and quick estimation of PED_h as a function of different values of few influencing attributes in order to perform benchmarking analysis or energy savings scenario analyses.

The paper is organized as follows: Section 2 analyses relevant works in the analysis of data from energy performance assessment; Section 3 describes the HEDEBAR methodology adopted to find a model for the characterization of heating energy demand; Section 4 shows the experimental results, which are then discussed in Section 5.

2. Related Work

Three main types of buildings *energy performance assessment* are commonly acknowledged [19]: *Energy benchmarking*, i.e., the comparison of Energy Performance Indicators (EPIs) of a building with a sample representative of similar buildings; *Energy rating*, i.e., the evaluation and classification of the building energy performance according to predefined criteria; and *Energy labeling*, i.e., the assignment of an energy performance class (or label) to the building, according to a scale of values defined for some relevant parameter (e.g., EUI, PED).

Energy rating can be implemented in the following ways: (i) *measured* (or *operational*) *rating*, based on real metering on-site [20] and (ii) *calculated rating*, based on ideal energy use. Measured rating is mostly used in the operation and maintenance phases of existing buildings [21]. Calculated rating is more suitable in the design phase of new buildings, in particular with the aid of Building Energy Simulation (BES) software like in the case of LEED and BREEAM rating systems [6,22]. Calculated rating is further divided into *asset rating* and *tailored rating*. While asset rating methods consider standard usage patterns and climatic conditions and can be shaped either to building designs or to existing buildings, tailored ratings consider actual conditions and usage patterns for the buildings under analysis.

Within the scientific context, several research activities have been carried out on buildings energy performance assessment, for: (i) predicting energy demand [7,10,23] and energy class [24], (ii) rating and benchmarking [25–28], (iii) individuating representative buildings for different classes of energy performance [29–31], (iv) characterizing the relationship between energy demand and relevant building features [32–34], and (v) improving existing methods, also using new model based on data mining algorithms like regression models, decision trees, neural networks, and clustering [24,32,35–38].

Several works have proposed a benchmarking of different types of buildings. Dall'O' et al. [25] analyse a real data set of energy certificates to assess the energy performance, to detect anomalies in the registered certificates and to quantify the energy retrofit potential in existing buildings. Chung et al. [26] developed a benchmarking process for energy efficiency of commercial buildings by means of Multiple Regression Analysis (MRA). Gao and Malkawi [29] use clustering to classify buildings according to multiple features, like physical properties, environmental conditions, occupancy. Lara et al. [30] adopt the cluster analysis to find out a few samples representative of about 60 buildings, in order to optimize the energy retrofit measures. Hong et al. [27] use an approach based on case-based reasoning, MRA, ANN and GA, to produce a methodology for operational rating with higher explanatory power and higher prediction accuracy at the same time. A parallel research effort by Acquaviva et al. [39] has been devoted to efficiently compute inter- and intra-building performance indicators on fine-grained thermal energy consumption data for a large set of buildings located in a major Italian city. Tso and Yau [37] compared the accuracy of linear regression, ANN, and decision tree in predicting average weekly electricity consumption during both summer and winter in Hong Kong. Koo et al. [7] use the finite element method to estimate the heating and cooling energy demand of buildings, using data about building envelope design. In [10] a decision tree is used to model the real consumption of residential buildings in order to predict the energy use of newly designed buildings. Melo et al. [24] use ANN to improve the accuracy of surrogate models for labeling purposes, based on simulations results. Khayatian et al. [35] tackle the problem of uniformity of criteria among different certificates, therefore they use ANNs to predict the heating energy demand and to validate a dataset of energy certificates.

The analysis of real data from EPC databases has been performed in various countries [11]. The authors in Fabbri et al. [40] discuss about the effects of EPBD Directive and Italian EPC system on the real estate market prospective. The study presented in Hjortling et al. [41] provides an energy consumption baseline for buildings in Sweden, using data from 186k energy performance certificates issued for commercial buildings and based on energy bills rather than on theoretical calculations. The paper shows that real energy consumption is often higher than the one stipulated by the building code. The methodology presented in Xiao et al. [42] exploits a cluster analysis of the energy consumption (EUI excluding District Heating) of office buildings in China, to study its statistical distribution characteristics. It was found that the distribution of energy consumption has quite different characteristics than in Japan and the US. Other analyses of EPCs aimed at defining the current energy consumption baseline of existing buildings in Greece and Spain are presented respectively in Dascalaki et al. [43] and Gangolells et al. [44].

Compared to the current state-of-the-art, this paper brings a contribution in the use of data mining techniques for the asset rating of buildings, both in methodological and analytical terms. From the *methodological* perspective, the paper proposes a novel approach to characterize the heating energy demand of buildings using multiple independent models for different building segments. From the *analytical* perspective, the proposed approach estimates the heating energy demand with reasonable errors, using a small set of building features and generating interpretable models that provide useful information about the most relevant features affecting energy demand.

3. Data Analysis Methodology

The HEDEBAR (*Heating Energy Demand Estimation for Building Asset Rating*) methodology estimates the *heating energy demand* of residential flats as a model of a few influencing features available within Energy Performance Certificates (EPCs).

HEDEBAR considers different building features that affect the energy demand. It is based on the hypothesis that the impact of each feature over the energy demand varies for different *segments* of values of the same energy demand. Hence, a *two-layer approach* has been defined to model this aspect. The logical components of HEDEBAR are represented in Figure 1 and they are briefly described below.

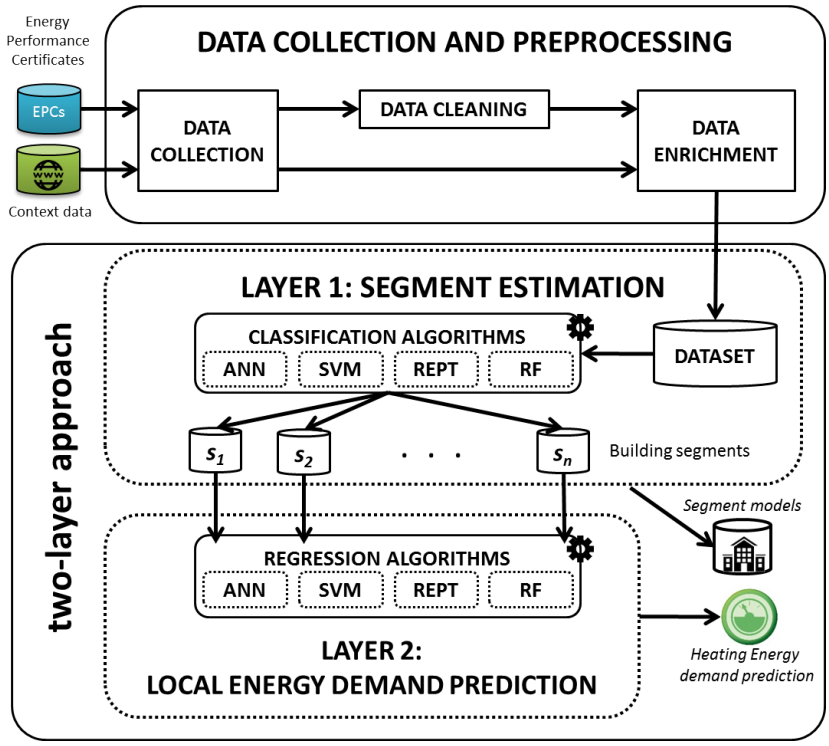


Figure 1. The proposed HEDEBAR methodology for automatic asset rating.

Data collection and preprocessing includes all the preliminary tasks necessary to provide the proper data set to the algorithms that operate in the later phases. Specifically, the *Data collection* component takes data from the energy certificates and other contextual information. *Data preprocessing* includes removing records with errors and missing values; discarding features that are useless to energy demand modeling; and enriching the resulting data set with contextual information not included in EPCs. These steps are better described in Section 3.2.

The *Segment estimation* is the first phase of the two-layer approach. Different classification algorithms have been trained during this step, to learn a classification model that properly assigns flats to different predefined segments of energy demand, considering only the selected features.

The *Local energy demand prediction* is the second phase of the two-layer approach. It uses regression algorithms to learn a regression model for estimating the *heating energy demand* considering only the selected features. An independent regression model for each segment of the first layer has been trained and tested.

During the two phases of the two-layer approach, the performance of each algorithm has been assessed in order to select the best one. When two or more algorithms have similar prediction performances, the one generating the most interpretable, i.e., human-readable model is preferred.

The two-layer approach provides a twofold output: the *classification and regression models* for the analyzed flats, useful to understand the features with the highest explanatory power with respect to the energy demand and to highlight the differences among the segments; the *heating energy demand prediction* for new flats.

3.1. Flat Characterization

The EPC includes the different features of a building affecting its energy performance as well as the variables used to quantify its energy demand. The feature selection process has been driven by previous experiences on EPCs datasets analysed by the authors [15,16] with the aim of using few input variables that are also easy to be collected. The following four main categories of input variables were identified for the purpose of the analysis: (i) *geometry*, (ii) *envelope*, (iii) *time*, and (iv) *system*. The categories are briefly described below, while Table 1 reports the relevant features for each of them.

Geometry. The variables in this category describe the different geometric features of the flat, which have an impact on its energy performance. The category includes variables such as average ceiling height, heat transfer surface and heated gross volume of the flat.

Envelope. The features in this category are related to the physical properties of the building (i.e., the thermal transmittance values of the opaque and transparent building envelope). In this category are also considered the dynamic characteristics of the building envelope through the variable q_{env} . This variable is expressed as an ordinal attribute that ranges from 1 to 5. The five quality classes are related to specific numerical ranges of time lag and decrement factor that can be extracted from a table provided in DM 26/6/2009 [45].

Time. This category includes time variables such as the building construction year.

System. This category includes features related to the heating system (i.e., the average system global efficiency for space heating). The average global efficiency of the heating system is calculated on the basis of the standard values of efficiency for each sub-system (generation, distribution, control, emission) according to UNI TS 11300-2 [14].

Among all the variables considered in this study, the *Primary Energy Demand for space heating* PED_h has been selected as the *target variable* of the analysis. PED_h (expressed in kWh/m^{2}y) is an energy related variable defined for benchmarking purposes. It is an estimation of the amount of real energy consumption of a flat in standard use conditions and it contributes to assign an energy class label to the flat. The PED_h value is estimated starting from the remaining *explanatory variables* included in Table 1 and can be used to compare different flats. In particular, similar pools of input variables proved to be robust enough for modeling in an effective way the building energy demand [15,16]. The PED_h value refers to the period of a heating season and it is normalized by the flat floor area. PED_h contributes to the evaluation of the overall Primary Energy Demand of flats (PED) together with the Primary Energy Demand for domestic hot water (PED_w). The heating energy demand is evaluated considering a building energy balance. The modelling of the building geometry considers real shapes and self or over shading of other buildings. The quasi steady-state calculation method is based on the monthly balance of heat losses (transmission and ventilation) and heat gains (solar and internal) evaluated in monthly average conditions. Transmission heat losses are estimated taking into account opaque and transparent surfaces and as well as the thermal bridging effect. In "Standard Rating", parametric values depending on floor area or heated net volume are taken into account when evaluating the ventilation rate and internal heat gains. The dynamic effects on the net heating energy demand are taken into account by introducing the dynamic parameters, utilization factor and an adjustment of the set-point temperature for intermittent heating/cooling or set-back. These parameters depend on the thermal inertia of the building, on the ratio of heat gains to heat losses and on the occupancy/system management schedules. The annual PED for space heating is calculated from the net energy demand through different system efficiencies (emission, control, distribution, generation) considering the thermal losses in the various sub-systems. For the heating season, the average system efficiency is defined as the ratio between the annual net energy and the annual PED for heating. The PED includes also the electrical energy demand of auxiliary systems.

Table 1. List of features selected to characterize and estimate the heating energy demand with the HEDEBAR asset rating methodology.

Category	Name	Symbol	Unit	Range
Explanatory variables				
Geometry	Floor area	A	m^2	$\mathbb{R}^+$
	Heat transfer surface	S	m^2	$\mathbb{R}^+$
	Average ceiling height	H	m	$\mathbb{R}^+$
	Gross Heated Volume	V	m^3	$\mathbb{R}^+$
	Aspect ratio	R	m^{-1}	$\mathbb{R}^+$
Envelope	Average U-value of vertical opaque envelope	U_o	W/m^2K	$\mathbb{R}^+$
	Average U-value of the windows	U_w	W/m^2K	$\mathbb{R}^+$
	Quality of building envelope	q_{env}	-	$\{1,2,3,4,5\} \subset \mathbb{N}$
Time	Construction year	y_c	y	$\mathbb{N}$
System	Average global efficiency for space heating	η_h	-	$[0,1] \subset \mathbb{R}$
Target variable				
Energy	Normalized primary energy demand for space heating	PED_h	kWh/m^2y	$\mathbb{R}^+$

3.2. Data Preprocessing

The whole raw data set gathered from EPCs usually includes many building features, represented through variables of different data types such as numeric (integer or real), nominal, textual, and boolean. However, some features could be not relevant for the subsequent data analysis and their inclusion in the features set would increase the complexity of the generated models. Most of the not selected variables are poorly related with the PED_h (e.g., textual descriptions, address of the flat) or include attributes with a high explanatory potential that are not so easy to be assessed without running a simulation in advance (e.g., heat losses for transmission, ventilation and infiltration). Moreover, data sets derived from energy certificates filled by auditors could contain imputation errors which can badly affect the quality of the extracted knowledge.

To address the above issues and to improve both accuracy and usefulness of the data analytics phase, HEDEBAR includes a preprocessing step. This step aims to (i) *clean* the original data collection to remove outliers and errors in data and (ii) *enrich* data with additional *contextual information* to cope with external environmental conditions that could differently affect the estimation of the PED_h value for each flat. These steps are better described below.

Data cleaning. The whole data set is firstly inspected based on the advice of domain experts to remove the less relevant features. In addition, on the selected input variables a data cleaning analysis was performed. The data cleaning phase is crucial in order to ensure the robustness of the analysis. In fact, EPCs datasets can be characterized low quality (in terms of attribute inconsistencies) [11]. However, the domain expertise in the energy and buildings field can prevent or at least limit inconsistency issues. According to [11] the consistency checks considered in this study are:

(i) Constraint rules for columns (e.g., area or volume cannot be negative); (ii) Domain expert analysis of values of the attributes (e.g., physical thresholds of system efficiency or thermal transmittance); (iii) Statistical checks (e.g., outlier detection though box plots).

Data enrichment. Data collected from the energy certificates are enriched with additional contextual information acquired from external data sources. To cope with external environmental conditions that could differently affect the estimation of the PED_h value for each flat, PED_h has been recalculated according to a reference standard climatic condition. In particular, all the EPCs issued in Piedmont

region are evaluated for both the standard climatic conditions of the actual city (in which the building is located), and the one of Turin. The PED_h considered as target variable in this study is then expressed for all flats as if they were located in Turin considering the same standard monthly outdoor temperature and solar radiation. Therefore, comparisons among flats can be done regardless of their actual location. However, if it is necessary to assess the performance of a flat in a city different from Turin, a data scaling based on standard Degree Days (DD) can be considered a valuable procedure. Specifically, to scale the estimated PED_h it is possible to multiply it for the ratio between the standard DD value of the city where the flat is located and the ones of Turin.

3.3. Two-Layer Approach for the Estimation of Heating Energy Demand

The HEDEBAR methodology makes use of the features from energy certificates as explanatory variables to predict the PED_h value of a flat.

The impact of each feature on the PED_h value can vary over different classes of energy efficiency. To cope for this aspect, distinct ranges of PED_h value, called *segments of energy demand* or simply *segments*, can be defined to partition the data set into groups of flats with more uniform energy efficiency. This segmentation allows HEDEBAR to analyze independently the different classes of flat energy demand (e.g., low, medium, and high).

The estimation of the PED_h value is structured in HEDEBAR as a *two-layer approach*, including two phases named *Segment estimation* and *Local energy demand prediction*. The two phases are applied in sequence to accurately predict the PED_h value of a flat:

- Firstly, the *Segment estimation* phase identifies the expected (discrete) *segment of energy demand* of the flat. The approach considers a set of reference segments of energy demand of a flat. This task has been modeled as a classification problem. A classifier is used to assign each flat to the corresponding (discrete) segment of energy demand based on its features.
- Then, the *Local energy demand prediction* phase predicts the (continuous) numeric value of PED_h for the flat, based on its features. This second task is formulated as a regression problem. A different regression model is trained in advance for each segment of energy demand.

Thus, in HEDEBAR a new flat (with unknown energy demand PED_h) is first classified into a segment of energy demand through the *Segment estimation* phase. Then, the PED_h value of the flat is estimated through the *Local energy demand prediction* phase, using the regression model assigned to that segment.

To generate the classification and regression models used in the two phases, the HEDEBAR system can easily integrate most classification and regression algorithms currently available in literature. To select the most appropriate algorithms, two complementary aspects were considered: (i) the ability of the algorithm to *accurately predict* the *segment of energy demand* and the PED_h value for a flat, and (ii) the *interpretability of the model* it generates. Based on these criteria, we selected four reference algorithms to be evaluated for integration in the two phases of HEDEBAR: *Artificial Neural Network* (ANN), *Support Vector Machine* (SVM) [46], *Reduced Error Pruning Tree* (REPT), and *Random Forest* (RF). ANN and SVM methods provided good performances for both classification and regression tasks in several applications. However, these methods generate non-interpretable models and are usually characterized by high computational cost for building the model. REPT and RF methods have good performances as well, but with overall lower computational costs. Moreover, REPT algorithm generates an interpretable model, which makes possible a better understanding of the relationship between the features and the energy demand. Finally, all the four algorithms have a good degree of robustness to outliers and missing values in the data set, even if in HEDEBAR these issues are handled in advance in the data preprocessing phases. The open source Rapid Miner v5.3.0 toolkit [47] and the statistical software R [48] have been used for the development of the classification and regression algorithms. The following paragraphs provide an overview of the main characteristics of four algorithms.

Artificial Neural Network (ANN). Inspired by the structure and behavior of biological neural networks, *Artificial Neural Networks* (ANNs) are often used to model complex relationships between input and output variables or to find patterns in data. An ANN consists of an interconnected group of nodes (neurons), organized in different layers, which receive inputs from other nodes and return as output a value computed as a function of suitably weighted inputs. A very popular type of ANN is the *feed-forward* neural network, where information moves through neurons only in forward direction, from the input to the output nodes.

The training of ANN is usually performed through *back-propagation* algorithm: the final outputs are compared with the correct values of training samples to compute the value of a predefined error-function. The error is then fed back through the network to adjust the weights of each connection in order to reduce the value of the error function. After repeating this process for a sufficiently large number of training cycles, the network usually converges to some state where the error of the calculations is small [49].

Support Vector Machine (SVM). Based on the work of Vladimir Vapnik in statistical learning theory [50], *Support Vector Machines* (SVMs) are a set of supervised learning methods, which can be used for classification or regression. A SVM model represents data samples as points in space, separated by a set of hyperplanes, so that the samples of the different categories are divided by a clear gap that is as wide as possible. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (*functional margin*), since, in general, the larger is the margin the lower is the generalization error of the classifier. When the samples are not linearly separable, *soft-margin* SVMs allow for classification errors during the training, to produce a more generic model for new data [51].

SVMs map samples into a higher-dimensional space, where presumably the separation is easier. However, the computational and storage requirements of SVMs increase rapidly with the number of training vectors and with the space dimension. To keep the computational load reasonable, SVMs use a kernel function K(x,y) that simplifies the computation of dot products in terms of the variables in the original space. The kernel function can be of different type such as linear, polynomial, sigmoid [49].

Reduced Error Pruning Tree (REPT). *Reduced Error Pruning Tree* (REPT) [52] is a fast decision tree learning algorithm that builds classification or regression trees using information gain or variance reduction as splitting criterion. More specifically, it generates multiple trees and it picks the best one, that will be considered as the representative. REPT uses *reduced error pruning* with *back fitting* method to prune the tree. At each iteration, a validation subset is used to estimate the Mean Square Error (MSE) on the predictions made by the tree. Starting at the leaves, each node is replaced with its most popular class and if the prediction accuracy is not affected then the change is kept.

Optimized for speed, REPT only sorts values of numeric attributes once at the beginning of the model preparation. Reduced error pruning has the advantage of simplicity and speed, moreover the representation of the data in form of a tree has the advantage, compared with other approaches, of being meaningful and easy to interpret.

Random Forest (RF). *Random Forest* is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [53]. The generalization error for forests converges almost surely to a limit as the number of trees in the forest becomes large. RF is based on *bagging*, a technique for reducing the variance of an estimated prediction function. Indeed, RF fits a number of decision tree classifiers on various sub-samples of the data set (and also on various subsets of features) and uses averaging to improve the predictive accuracy and to control over-fitting. The resulting model is a voting model of all the random trees in the forest.

4. Case Study

In this section we validate the effectiveness and the usability of the proposed HEDEBAR methodology focusing on the following aspects: (i) the ability to correctly estimate the segment of energy demand for each flat, and (ii) the ability to accurately predict the PED_h value for each flat. The experimental analysis also addresses (iii) the selection of the classification and regression algorithms integrated in the two layers of the system, (iv) the comparison with a single layer approach in terms of prediction error and overall execution time, (v) the impact of the system configuration parameters, and (vi) the explanation of the main variables that determine the membership of flats into segments and their PED_h values.

We experimentally evaluated HEDEBAR on a real data collection of EPCs issued in 2013 for buildings located in the Piedmont region, North West of Italy. The data set includes approximately 90,000 energy certificates, of flats located across the 8 provinces of the Piedmont region.

4.1. Characterization of Flat Segments

As explained in the methodology section, the data set has been partitioned into different segments according to the values of variable PED_h with the aim of grouping together flats with similar energy efficiency.

Specifically, three reference segments have been considered representing respectively *low energy demand flats* (segment s_1), *high energy demand flats* (s_2), and *very high energy demand flats* (s_3). Data set splitting into segments has been done considering also the reference value range of PED_h specified in [15,16]. Segment s_1 includes flats with PED_h values between 0 and 100 kWh/m^{2}y, while flats in segment s_2 have 100 kWh/m^{2}y $\leq PED_h \leq$ 300 kWh/m^{2}y, and in segment s_3 $PED_h \geq$ 300 kWh/m^{2}y.

The three segments result into sets with the following cardinalities. The larger segment is s_2 including 39,003 flats, followed by s_1 with 25,930 flats, and the s_3 with 21,176 flats.

The dataset has been split into three segments to identify representative groups of energy performance certificates representing flats with similar performances. Specifically, a group represents flats with low energy demand, the second includes flats with medium-high energy demand, while the last one includes flats with very high energy demand. The three segments also allow guaranteeing a significant number of flats in each group together with a variable distribution for each feature under analysis. A number of segment higher than three should lead to very small groups of energy performance certificates with a limited data variability for each variable. In this case an estimation model for a segment should not be general (i.e., data overfitting). A small number of segments should lead to the definition of complex estimation models of heating energy demand. In this case derived models could not be easily understood and quickly exploited by a domain expert.

Box plots in Figure 2 show the distribution for some interesting variables (i.e., average U-value of vertical transparent envelope, average global efficiency for space heating, construction year, and aspect ratio) separately for each segment under analysis. In general, all segments present a good variability range for each variable under analysis. Specifically, segment s_1 includes a set of residential flats characterized by a low energy demand. In fact, flats in this group are characterized by the lowest values of U_w (median 2.11, IQR $[1.75, 2.76]$), U_o (median 0.45, IQR $[0.33, 0.67]$) and R (median 0.6, IQR $[0.4, 0.7]$); and the highest values of η_h (median 0.81, IQR $[0.73, 0.87]$) and y_c (median 2004, IQR $[1970, 2009]$). On the other hand, segment s_3 includes flats characterized by a very high energy demand, represented by the highest values of U_w (median 3.66, IQR $[2.80, 4.62]$), U_o (median 0.98, IQR $[0.83, 1.04]$) and R (median 0.9, IQR $[0.7, 1.0]$); and the lowest values of η_h (median 0.68 range $[0.60, 0.73]$) and of y_c (median 1962, IQR $[1940, 1973]$. Finally, segment s_2 is characterized by median values and IQRs of the five variables that lie between those of the two previous segments.

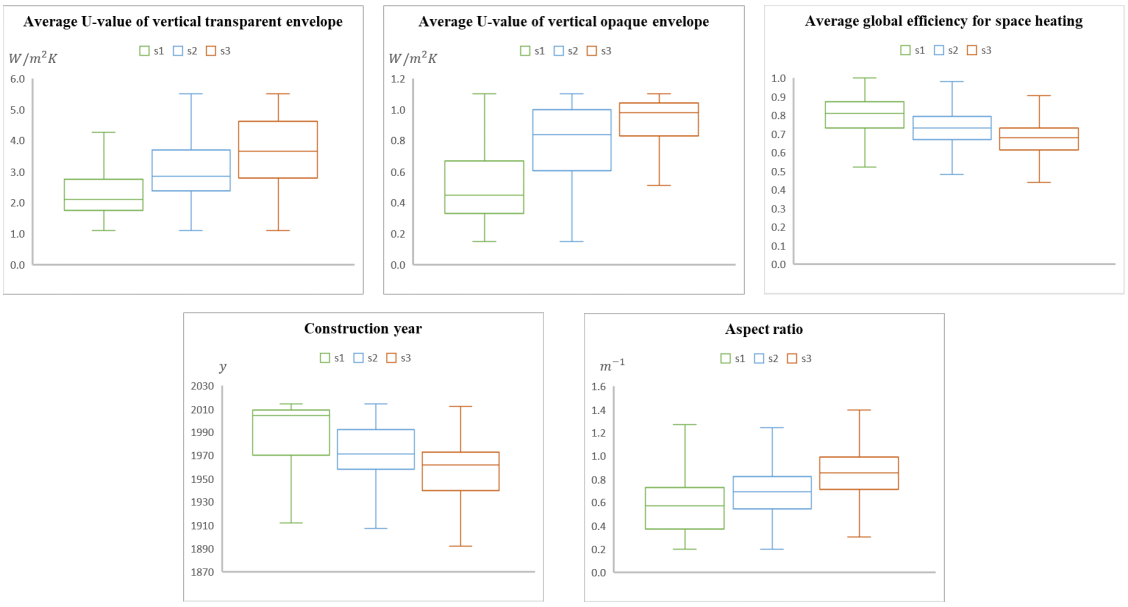


Figure 2. Box plots of the values of 5 input variables evaluated for each of the three different segments of energy demand.

Figure 3 shows the distribution of the certificates across the 8 Piedmont provinces, separately for each segment. The three charts are quite similar to each other, demonstrating that the geographical distribution is very similar across the three segments.

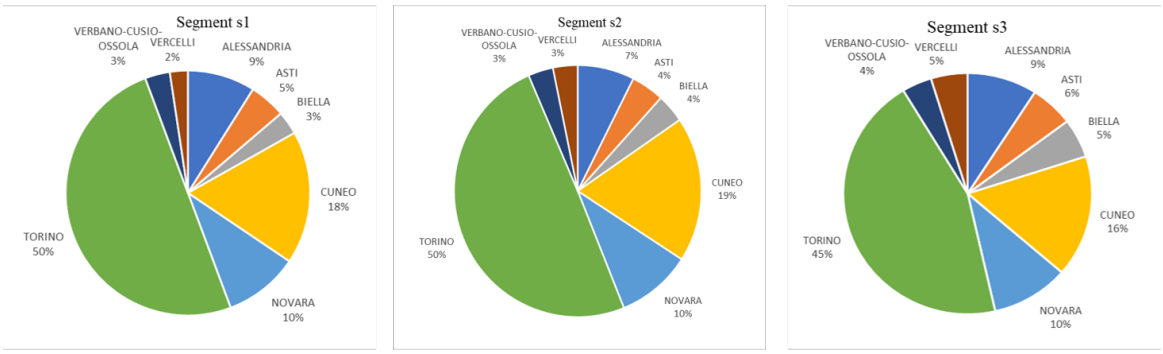


Figure 3. Distribution of the buildings across the 8 provinces of the Piedmont region for each of the three different segments of energy demand.

4.2. Segment Estimation

The classification task aims at assigning each new flat into the correct segment of energy demand. The classes of the classification task are the three segments presented in Section 4.1, identified by the nominal labels s_1, s_2, and s_3. All the four classification algorithms integrated in HEDEBAR (i.e., ANN, REPT, RF and SVM) have been experimentally evaluated for the classification of flats. The algorithm providing the classification model with the best classification performance has been selected as reference for this phase.

To validate the results of the classification process four established performance measures [54] have been considered. The overall quality of the classification model is evaluated in terms of *accuracy*. This measure counts the total number of flats correctly assigned to their corresponding segment. However, the unbalanced distribution of flats in the three segments could lead to a biased value of accuracy, as it could be mostly influenced by bigger segments. Therefore, other measures have been also used for a more accurate evaluation of the classification model. Per-class classifier predictions were evaluated according to *precision*, *recall*, and *F1-measure*. *Precision*(s_i) indicates the percentage of flats

that are correctly revealed as in segment s_i. $Recall(s_i)$ indicates the number of flats assigned to segment s_i with respect to the total number of flats actually in s_i. The $F1\text{-}measure(s_i)$, which is computed as the harmonic average of $Precision(s_i)$ and $Recall(s_i)$, quantitatively estimates the balancing between $Recall(s_i)$ and $Precision(s_i)$. In the experiment evaluation, we computed the precision, recall, and F1-measure values for each class label corresponding to each of the three segments.

A good trade-off between recall and precision is needed to properly predict the PED_h values for a new flat. On the one side, high precision values on most (all) segments are crucial to foster an accurate prediction of the PED_h values in the subsequent regression task. Indeed, the correct classification of a flat into the corresponding segment facilitates the subsequent prediction of the PED_h value for the flat. In fact, this prediction is performed through a model trained using data of flats with similar energy performance. A low $Precision(s_i)$ value indicates that many flats were mistakenly classified into segment s_i. This would result in erroneous predictions of PED_h values in the second step. On the other hand, achieving high recall values on most segments is desirable as well. A low $Recall(s_i)$ indicates that few flats of segment s_i are correctly classified into s_i, and they have been wrongly assigned to a segment other than s_1. This wrong assignment would result into an erroneous predictions of PED_h values due to the selection of a less appropriate prediction model in the second step.

Table 2 reports the results achieved by the four classification algorithms integrated into HEDEBAR. It shows the accuracy on the overall data set as well as precision, recall, and F1-measure for the three segments.

Table 2. Overall classification accuracy and precision, recall and F1-measure for each segment of ANN, REPT, RF and SVM algorithms.

	ANN	**REPT**	**RF**	**SVM**
Overall				
Accuracy (%)	67.51	82.03	85.67	67.24
Segment s_1				
Precision (%)	77.71	87.70	90.52	82.49
Recall (%)	70.03	83.84	87.27	61.97
F1-measure (%)	73.67	85.73	88.87	70.77
Segment s_2				
Precision (%)	62.11	80.40	82.65	60.68
Recall (%)	75.54	80.56	85.49	81.74
F1-measure (%)	68.17	80.48	84.05	69.56
Segment s_3				
Precision (%)	68.65	78.60	83.58	70.62
Recall (%)	49.62	82.53	81.96	46.98
F1-measure (%)	57.60	74.93	82.76	56.42

The RF classifier provides the highest accuracy value (85.67%) followed by REPT (82.03%), ANN (67.51%) and SVM (67.24%). Moreover, RF achieves also the best F1-measure on all segments (88.87%, 84.05%, and 82.76% in segments s_1, s_2 and s_3 respectively). More in detail, RF obtains the highest precision value for all segments (90%, 82.65%, and 83.58% for segments s_1, s_2, and s_3 respectively). RF also provides the highest recall values for two segments (87.27% and 85.49% for segments s_1 and s_2 respectively), while the recall obtained on segment s_3 (81.96%) is very close to the value provided by algorithm REPT (82.53%), which is the highest recall value over the four algorithms. Since the RF classifier achieves the highest values for almost all performance parameters, we chose it as reference algorithm for creating the model which classifies a new flat into the corresponding segment.

REPT is the second best algorithm for almost all performance parameters, providing accuracy, precision and recall values lower than those of RF, but still more than acceptable. An additional key point of REPT is the fact that this algorithm builds an interpretable classification model. This model is a decision tree from which human-readable classification rules can be extracted. Thus, domain experts

can use the model not only to automatically classify a flat into the corresponding segment but also to analyze the most relevant properties that characterize each segment as well as to understand why a flat has been classified into a segment (see Section 4.5.1).

The SVM and ANN algorithms provide the worst values for all performance parameters, which are significantly lower than those obtained with RF and REPT algorithms.

Therefore, according with the experimental evaluation we decided to include two different classification models into the *Segment estimation* layer of the HEDEBAR framework. The RF classifier is used to automatically label a new flat with the corresponding segment. Based on the assigned segment, the proper regression model is selected in the subsequent layer (*Local energy demand prediction*) to predict the PED_h value for the flat. Instead, the REPT model is used to provide domain experts with a qualitative analysis of the impact of variables characterizing flats on the primary heating energy demand. This aspect is further discussed in Section 4.5.1.

4.3. Local Energy Demand Prediction (PED_h)

The regression task aims at estimating the value of PED_h for a flat. In HEDEBAR a different regression model for PED_h prediction is created for each of three segments s_1, s_2, and s_3. The ANN, REPT, RF and SVM algorithms have been experimentally evaluated for the creation of the regression model for each segment.

Table 3 displays the mean prediction errors of the four algorithms in predicting PED_h for each segment as well as the mean errors averaged over the three segments. The *prediction error* is the difference between the real value and the predicted value of PED_h. Three different measures of prediction error, among those commonly used in literature, have been calculated: (i) *Mean Absolute Error* (MAE) is the mean of all the absolute values of the errors obtained with the test samples; (ii) *Mean Absolute Percentage Error* (MAPE) expresses the mean absolute error in percentage terms; (iii) *Root Mean Square Error* (RMSE) is the square root of the mean of the square of all the errors obtained with the test samples. While MAE refers only to the mean value of the distribution of absolute errors, RMSE is affected also by the standard deviation of such distribution. Compared to MAE, RMSE amplifies and severely punishes large errors.

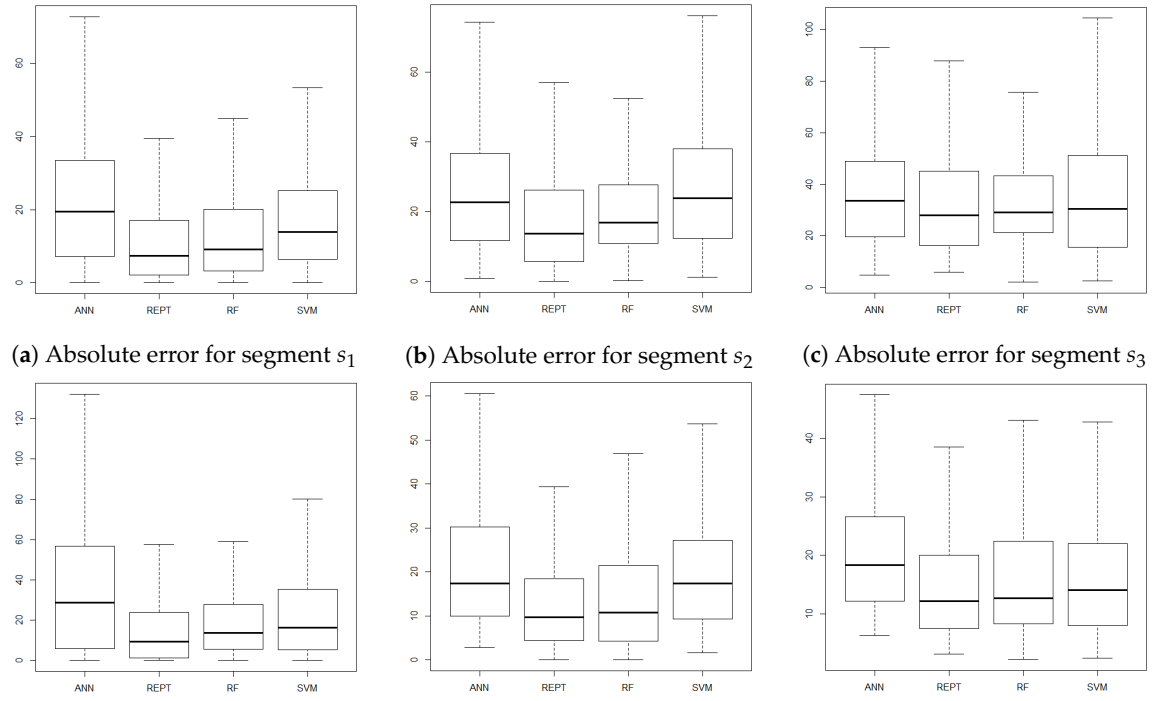
Table 3. Errors in predicting PED_h for ANN, REPT, RF, and SVM algorithms and for each flat segment.

	ANN	**REPT**	**RF**	**SVM**
Overall				
RMSE (kWh/m^2)	39.85	33.12	33.83	38.40
MAE (kWh/m^2)	29.67	22.21	22.35	27.41
MAPE (%)	27.02	16.64	16.89	21.52
Segment s_1				
RMSE (kWh/m^2)	30.99	21.99	22.16	28.95
MAE (kWh/m^2)	23.04	13.45	13.88	18.83
MAPE (%)	40.76	20.25	20.47	27.32
Segment s_2				
RMSE (kWh/m^2)	37.80	29.72	30.87	37.03
MAE (kWh/m^2)	28.23	20.57	21.52	28.02
MAPE (%)	22.33	14.75	15.62	20.37
Segment s_3				
RMSE (kWh/m^2)	49.76	47.69	49.84	50.31
MAE (kWh/m^2)	38.78	36.26	37.53	37.76
MAPE (%)	20.87	15.90	17.18	17.19

The REPT algorithm produces the overall lowest error values for the three measures (MAPE = 16.64%, RMSE = 33.12 kWh/m^{2}y, MAE = 22.21 kWh/m^{2}y) and it has also the best performance in each segment. In relative terms, REPT performs better in segments s_2 and s_3, where MAPE is 14.75%,

and 15.90% respectively, while it has a substantially lower performance in segment s_1, where MAPE = 20.25%. The second best algorithm is RF, with an overall MAPE of 16.89%, while SVM and ANN provide higher error values (MAPE = 21.52% and MAPE = 27.02% respectively). Therefore, the REPT algorithm has been selected for local energy demand prediction, in order to better characterize groups of flats with similar features.

Figure 4 analyses more in depth the distribution of prediction errors, by reporting the box plots for *absolute error* and *percentage error* of the four algorithms over the three segments. The difference between REPT and the other algorithms is clear especially in segments s_1 and s_2.



(**a**) Absolute error for segment s_1 (**b**) Absolute error for segment s_2 (**c**) Absolute error for segment s_3

(**d**) Percentage error for segment s_1 (**e**) Percentage error for segment s_2 (**f**) Percentage error for segment s_3

Figure 4. Box plots of absolute error and percentage error of estimation of energy demand for each algorithm and for the three different flat segments.

4.4. Performance Comparison with a Single Layer Approach for PED_h Prediction

In this section we compare the performance in the prediction of the PED_h value between the *two-layer approach* used in HEDEBAR and a *single layer approach*. This latter approach exploits a unique regression model for all three segments, instead of building different models tailored to each segment. The ANN, REPT, RF and SVM algorithms have been evaluated to build the regression model for PED_h prediction with the single layer approach. The configuration setting for the single layer approach is discussed in Section 4.6.

Results for the two-layer and single layer approaches are reported in Tables 3 and 4, respectively. The experimental evaluation showed that, as for the two-layer approach, also for the single layer approach the best performance for PED_h prediction is obtained using the REPT algorithm. However, the REPT algorithm applied to the overall data set provides a model with MAPE value equal to 21.26% (see Table 4). Instead, using the two-layer approach the REPT models tailored to each segment result into a significantly lower overall MAPE value, equal to 16.64% (see Table 3). Also the RMSE and MAE values are significantly higher with the single layer approach (respectively, 37.37 kWh/m^2 and 26.10 kWh/m^2) than with the two-layer approach (respectively, 33.12 kWh/m^2 and 22.21 kWh/m^2). These results demonstrate the suitability of the two-layer approach used in HEDEBAR. In fact, the

segmentation of the entire data set into groups of flats with similar energy demand allows to build differentiated models, which can more precisely predict the PED_h value for a flat in the segment.

Table 4. Errors in predicting PED_h for ANN, REPT, RF and SVM algorithms using a single step regression.

	ANN	**REPT**	**RF**	**SVM**
RMSE (kWh/m^2)	45.33	37.37	38.03	42.65
MAE (kWh/m^2)	30.01	26.10	26.36	28.34
MAPE (%)	27.46	21.26	21.53	23.67

4.5. Interpretation of the Energy Demand Estimation Models

This section provides a qualitative analysis of the impact of explanatory variables (building features) on the dependent variable, (heating energy demand). The analysis makes use of the REPT model, which has the advantage of providing interpretable decision trees. To better understand how the REPT algorithm models the relationship between input variables and the heating energy demand, we illustrate the first levels of the obtained decision trees.

4.5.1. Segment Estimation Model

The descriptive power of the REPT model comes from its capacity of putting in evidence the features that mostly affect the energy demand, according to the analyzed certification system.

The REPT model is represented by a tree graph, made of nodes and leaves connected by edges. In the REPT model built in HEDEBAR for segment estimation, each path of the tree includes a subset of building features. The leaf node of a path represents the predicted class label, corresponding to the energy demand segment s_1, s_2 or s_3 in this study. Therefore, each tree path includes a subset of features describing the buildings in one of the three segments.

A common way to build such trees is based on a recursive partitioning method. It consists in a forward step-wise approach where at each node the best split (according to input split variable, and the split value) is automatically evaluated by the algorithm for maximizing homogeneity in its child nodes. In this way the selection of split variables and split values consists in a data-driven process that does not require a manual selection by the analyst. As an example, the node including the *construction year* feature (y_c) can include the value 2007 as splitting value. The two outgoing edges for the node are associated to two distinct sets of values for y_c such as for example $y_c < 2007$ and $y_c \geq 2007$. Thus, each path includes a subset of variables, together with their corresponding ranges of values, describing the buildings associated with the segment label appearing in the leaf node of the path. For the classification of a new flat, the tree path composed of all the edges with splitting rules satisfying the features of the flat is selected. The segment label appearing in the leaf node of the path is used to estimate the segment of energy demand for the flat.

The first four levels of the REPT model are illustrated in Figure 5 (please refer to Table 1 for the interpretation of input variable symbols). It is possible to observe that the *average U-value of vertical opaque envelope* parameter (U_o) is the one mostly affecting the energy demand. Also the *aspect ratio* (R) and the *construction year* (y_c) appear at the first three levels of the tree. *Average U-value of the windows* (U_w) and *average global efficiency for space heating* (η_h) appear only at the fourth level. In general, the splits closest to the root node are the most important ones. This is the reason why only the upper portion of the classification tree is shown in Figure 5.

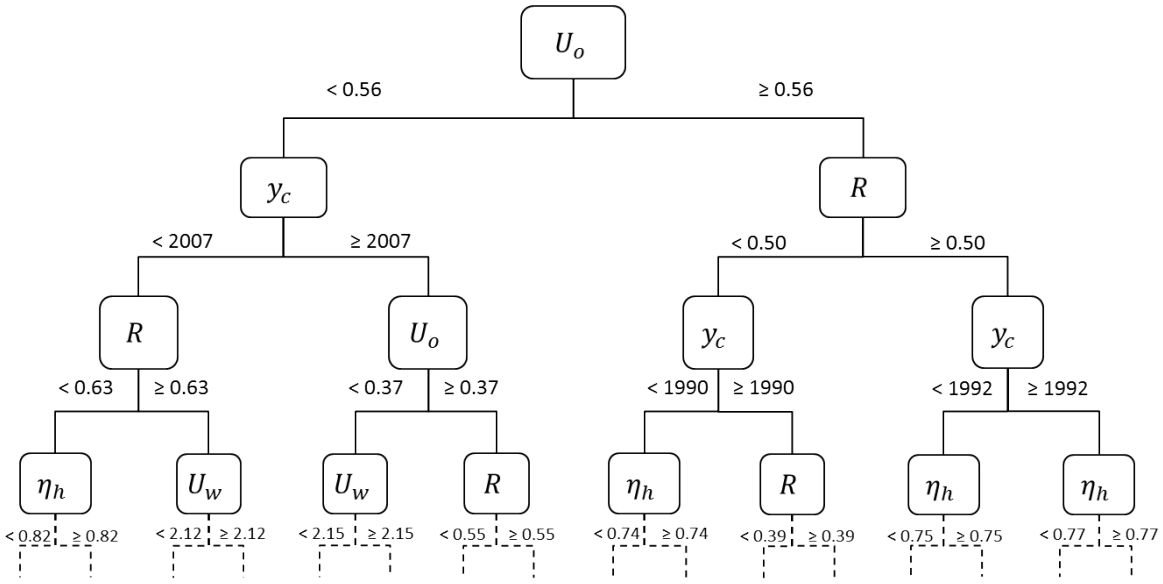


Figure 5. REPT model of the classification phase. The first four levels of the tree are illustrated and, for each path, the histogram illustrates the number of leaves assigned to each segment.

To further facilitate the interpretation of the tree model and to highlight the characteristics of each segment, the classification rules that summarize the main paths of the tree were extracted. The model developed for segment estimation has an overall size of 342 nodes with a maximum depth of 20 levels. Identify the most significant paths of the tree means to extract from the set of decision rules the ones that involve a significant number of records and reach high values of accuracy. These rules bring out the most representative building properties of each segment together with their ranges of values. Rules are extracted by traversing tree paths and they are structured in two parts: (i) the *rule antecedent* includes the buildings features and the corresponding ranges of values; (ii) the *rule consequent* includes the energy demand segment associated to flats that satisfy the conditions of the rule antecedent. Table 5 resumes the subset of rules selected as reference example from the REPT model. Specifically, for each segment we selected the rules with the highest classification accuracy among those that classify at least 500 flats. For the selected paths, the classification accuracy, i.e., the percentage of flats classified into the correct segment, ranges from 74.7% to 93.7%.

Table 5. Main rules of the REPT model for classification. For each row, intervals are specified only for the variables used by the corresponding rule. The last column contains the segment assigned by the rule.

		Rule	**Antecedent**			**Rule Consequent**
U_o	y_c	R	U_w	η_h	q_{env}	**Segment**
$[0, 0.37[$	$[2007, +\infty[$		$[0, 2.15[$			$\Rightarrow s_1$
$[0.56, +\infty[$	$[1992, +\infty[$	$[0.5, 0.68[$		$[0, 0.77[$		$\Rightarrow s_2$
$[0.78, +\infty[$	$]-\infty, 1991]$	$[0.63, 0.98[$	$[3.41, +\infty[$	$[0, 0.75[$	$[2, 5]$	$\Rightarrow s_3$

Rules like those in Table 5 are an important source of information about the classification model. Therefore, by examining these decision rules, the significant factors influencing PED_h can be identified also by a non-expert user and it is possible to roughly estimate the segment of a new flat.

For instance, the rule for segment s_1 is based on the average U-values of vertical opaque envelope (U_o) and of the windows (U_w) and on the construction year (y_c). More specifically, the rule states that, if $U_o < 0.37$ W/m^{2}K and $U_w < 2.15$ W/m^{2}K, the building envelope guarantees a very high level of thermal insulation and low heat dissipation. Moreover, flats that satisfy this rule were built with

construction standards adopted from 2007 onwards, thus guaranteeing an overall energy efficiency that is classified into segment s_1.

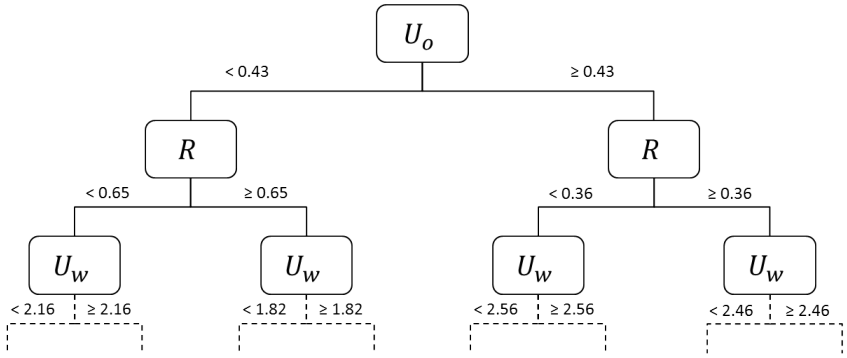
The rule for segment s_2 includes also the aspect ratio (R) and the average global efficiency for space heating (η_h). This rule shows that, for high energy demand flats, R has intermediate values, while the η_h is always lower than 0.77. The average U-value of vertical opaque envelope (U_o) has a minimum value of 0.56 W/m^{2}K, which is higher than the maximum value used in the previous rule of s_1 (0.37 W/m^{2}K), thus implying always a higher thermal transmittance. Moreover, the rule includes high energy demand flats constructed since 1992, i.e., the minimum construction year for this rule is 15 years lower than the one for the previous rule (2007).

The rule selected for segment s_3 has very high values of aspect ratio (R), starting from a minimum of 0.63 m^{-1} which is almost equal to the maximum value for s_2 (0.68 m^{-1}). Additional negative factors are represented by the high lower bounds for U-values (U_o, U_w) and the construction year (y_c) always before 1991.

4.5.2. Local Energy Demand Prediction Models

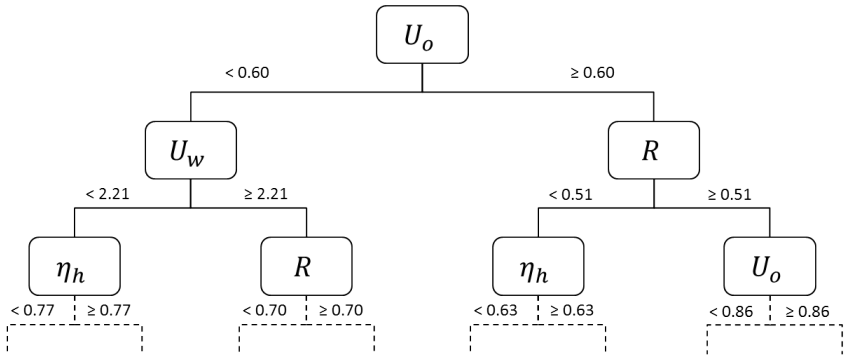
Figure 6 depicts the first three levels of the REPT regression models of *Local energy demand estimation* for the three flat segments. Variables of splitting rules associated to the tree nodes are almost the same of the classification model represented in Figure 5, however their importance vary according to the segment. The tree for segment s_1 has a single variable for each level, i.e., *U-value of vertical opaque envelope* (U_o) at the first, *aspect ratio* (R) at the second, and *U-value of the windows* (U_w) at the third, thus providing a simple and easily interpretable model. In segment s_2 the *average global efficiency for space heating* (η_h) has a higher importance than in s_1, as it appears at the third level of the tree. The same variable appears in most of the rules of the same level in segment s_3. Here *average U-value of the windows* (U_w) is considered only for the most efficient flats (with $U_o < 0.76$ W/m^{2}K and $R < 0.89$ m^{-1}), while for those with higher energy demand, the *average global efficiency for space heating* (η_h) becomes more significant.

The splitting value of *average U-value of vertical opaque envelope* (U_o) increases from segment s_1 to segment s_3, meaning that flats belonging to the first segment are characterized by higher thermal insulated walls.

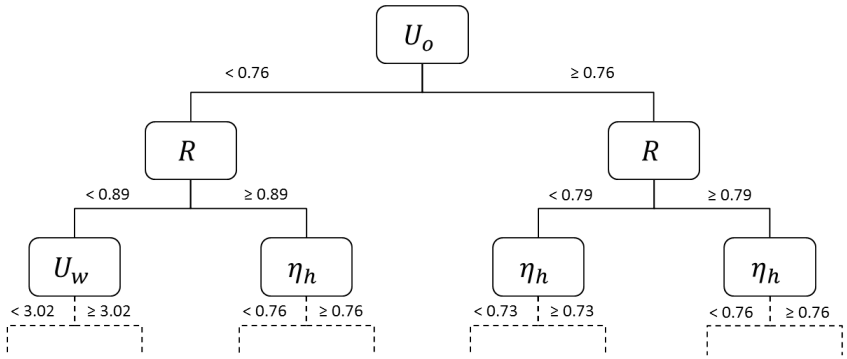


(**a**) Segment s_1

Figure 6. *Cont.*



(**b**) Segment s_2



(**c**) Segment s_3

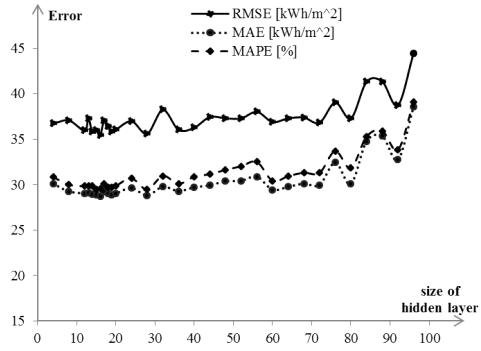
Figure 6. REPT models for each of the 3 flat segments.

4.6. Parameter Tuning of Algorithms

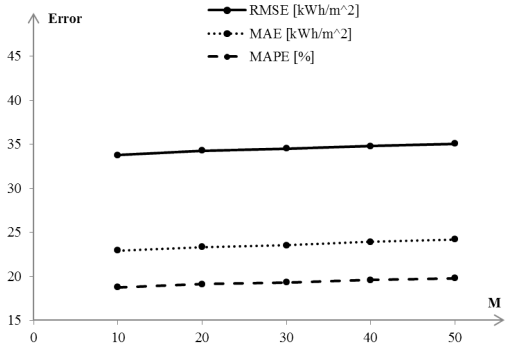
This section describes how the main parameters of the four algorithms considered in this study were tuned in order to reach the lowest values of prediction error both in the *Segment estimation* and *Local energy demand prediction* phases in the HEDEBAR framework. The same tuning procedure has been used also for the configuration of the single layer approach considered for performance comparison and described in Section 4.4.

For both phases, the prediction error was assessed using the k-fold cross-validation method, with $k = 10$. Therefore, the input dataset for the target phase has been split into k subsets of the same size. In turn, 1 subset is used for testing and the remaining $k - 1$ are used for training. Hence, k independent training and test iterations are performed. For each iteration, the training set is used by the four algorithms to generate the classification or regression models, according to target phase in the HEDEBAR framework. Then, the test set is used to evaluate the capacity of each classification and regression model to predict respectively the segment of energy demand and the PED_h value of new flats. The overall error value after the k iterations is computed as the mean of the errors of the k tests.

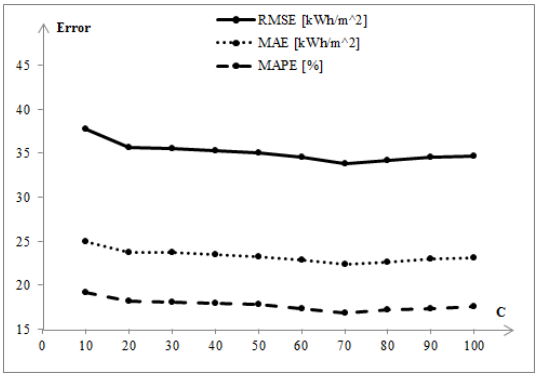
The procedure for tuning the optimal configuration for each of the four algorithms used in HEDEBAR produced similar values of parameter settings for the creation of the classification and regression models. These parameter settings turned out to be the optimal configuration even for the single layer approach. As an example, this section describes the results of parameters tuning for the creation of the regression model used in the *Local energy demand prediction* phase. The parameter tuning procedure is aimed at minimizing the values of the prediction errors MAPE, MAE, and RMSE (Figure 7).



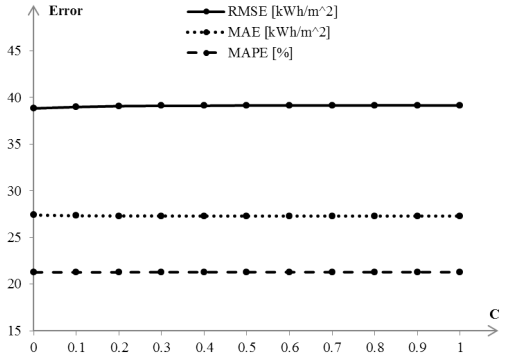
(**a**) ANN algorithm with respect to the size of the hidden layer.



(**b**) REPT algorithm with respect to the minimum number of instances per leaf *M*.



(**c**) RF algorithm with respect to the number of trees.



(**d**) SVM algorithm with respect to the complexity constant *C*.

Figure 7. Overall Local energy demand prediction errors of the algorithms for different values of their parameters.

For the ANN algorithm, a single hidden layer of variable size was considered, since using more than one layer did not provide any significant improvement of accuracy. Some common rules of thumb for the size of the hidden layer in the ANN are suggested by different works like [55], where the number of neurons are related to the number of input and output variables. Overall, the size of the hidden layer should be high enough to let the ANN model the problem correctly, but also low enough to ensure generalization. An increasing number of neurons was used during the tests, ranging in the interval $[4, 100]$ until the prediction error starts to grow due to over-fitting. The other parameters of the ANN are: $learning\_rate = 0.3$, $training\_cycles = 10^3$, $\epsilon = 1 \times 10^{-5}$. The values of RMSE, MAE and MAPE for different sizes of the hidden layer are reported in Figure 7a. 16 neurons for the hidden layer provide the lowest values of the three errors.

In the REPT algorithm, the dimension of the pruning subset was set to one third of the training set, hence with three folds in the algorithm ($N = 3$). No maximum tree depth has been set instead. The *information gain* was used as splitting criterion. The REPT algorithm was tuned by varying the minimum number of instances per leaf ($M \in \{10, 20, 30, 40, 50\}$). The values of RMSE, MAE and MAPE are reported in Figure 7b. The three error measures slightly, yet constantly, increase together with M. Therefore M was set equal to 10.

In the RF algorithm, the previous settings of REPT was used for all the decision trees. The variation of prediction error was assessed with respect to the number of trees I in the range [10, 100]. The values of RMSE, MAE and MAPE are reported in Figure 7c. $I = 70$ provides the lowest error values.

For SVM regression, a linear kernel function was considered and the variation of prediction errors, with respect to the complexity constant C, was assessed. This variable is used to set a degree of tolerance for misclassification of training samples. A too large value of complexity constant can lead to

over-fitting, while too small values may result in over-generalization. Values for C have been selected in the range $[0, 10]$. The other parameter settings of the SVM are: $max\_iterations = 10^4$, convergence $\epsilon = 1 \times 10^{-3}$. The values of RMSE, MAE and MAPE are reported in Figure 7d. The trends of the three error measures are nearly constant with a slightly lower value of RMSE for $C = 0$.

5. Discussion and Conclusions

In this paper, the HEDEBAR methodology for the automatic asset rating of flats energy efficiency has been described. We recall that the analysis has been possible thanks to the availability of open data of Energy Performance Certificates. HEDEBAR proposes a two-layer approach to compute the ideal *Primary Energy Demand for space heating* (PED_h) of flats according to the certification scheme used to issue their EPCs. In this section we discuss the results obtained through HEDEBAR, addressing the results achieved using the proposed two-layer approach, and the interpretation and the possible exploitation of the extracted knowledge.

Accurate estimation of the flat energy demand with a reduced features set. Experimental results demonstrated the ability of the HEDEBAR methodology to estimate the PED_h value for a flat. PED_h is not the actual energy consumption of a flat, but its primary energy demand calculated in standard conditions. It is a significant parameter for the comparison of flats based on their features. The estimated values of PED_h are precise enough to provide a dependable assessment of flat energy efficiency for different values of the features characterizing flats.

From a methodological perspective, the experimental evaluation demonstrated that the two-layer approach used in HEDEBAR performs significantly better than a single layer algorithm in estimating the PED_h (MAPE values are respectively 16.64% and 29.82%). Therefore the segmentation of the initial data collection into different groups of flats with similar energy demand allows to produce differentiated models, which fit better the specific features of the respective segments.

The predictive performance of the HEDEBAR methodology is similar to the one of Khayatian et al. [35], where ANNs are used to predict the PED_h value, using EPCs related to the Lombardy region. Indeed, even if the experimental evaluation has been conducted on different datasets, HEDEBAR and the approach in [35] provide comparable results (MAPE equal to 16.64% HEDEBAR and to 14.44% in [35]). However, differently from [35], HEDEBAR estimates the value of PED_h in two steps using the REPT algorithm, which provides an interpretable model.

Modular approach able to integrate various algorithms and applicable to EPCs from other certification schemes. The HEDEBAR approach can make use of various classification and regression algorithms and can be used also to analyze data of EPCs issued according to other certification schemes.

The performed experimentation puts in evidence the algorithms with the best performances among those which were tested. In the *Segment estimation* phase, RF algorithm has the highest classification accuracy, while, in the *Local energy demand prediction* phase, REPT algorithm has the lowest error values in predicting PED_h. REPT also has a good classification accuracy. Therefore, RF in the first and REPT in the second phase turned out to be the most suitable combination of algorithms for the estimation of PED_h from the variables included in the EPC data set.

Interpretation of the energy demand estimation models. A key advantage of HEDEBAR is the use of REPT algorithm, whose decision tree models make results understandable and exploitable also for non-domain experts. Useful information can be obtained from this model as it helps to discover in a straightforward way energy patterns among large dataset. The algorithm automatically selects the different attributes for generating split rules and the ones closest to the root node can be assumed as the most influencing attributes. Therefore, the performance improvement brought by the two-layer approach, especially to the REPT algorithm, provides the HEDEBAR methodology with both a good estimation precision and a set of interpretable models of energy demand. Resulting models pointed out the most relevant features according to the considered rating system.

In the *Segment estimation* layer, 5 features out of 10 (*average U-values of opaque envelope and of the windows*, *aspect ratio*, *construction year*, and *average global efficiency for space heating*) appear in the first four levels of the decision tree and can be considered as the most relevant ones of the model. Indeed, they were preferred to other variables for splitting the initial flat set since they generate more homogeneous subsets in terms of PED_h value, thus allowing the overall model to reach a more accurate segmentation of the flat set. The characteristics of the three segments of energy demand are also summarized by means of short *decision rules*, which bring out the most representative building properties and their ranges of values for each segment. With a view to improving the efficiency of a flat, the model makes possible to individuate the features that mostly cause its membership to a specific energy demand segment. A proper change of their values, when possible (e.g., by means of targeted refurbishment actions), can substantially increase the energy efficiency of the flat. For some flats, bringing the values of few features within the appropriate ranges causes their reassignment to a lower segment.

In the *Local energy demand prediction* layer, 4 features out of 10 appear in the first three levels of the three decision tree models (the same as in Segment estimation except *construction year*). The differentiated analysis highlighted the main features impacting on PED_h for different segments of energy demand. In this case, the *U-value of vertical opaque envelope* (U_o) has demonstrated to be one of the most important variables for all segments. Indeed U_o is at the first level of all the three REPT models, with increasing splitting values from s_1 to s_3. The aspect ratio (R) is also a significant variable, as it appears in the second level of all the three REPT models. The *average U-value of windows* (U_w) is more important for low levels of energy demand (segment s_1), where the contribution of heat loss through windows can make the difference. On the other hand, the relevance of the *overall efficiency of the heating system* (η_h) is evident only for *high* and *very high* energy demand flats (segments s_2 and s_3).

Possible exploitation of HEDEBAR findings. Energy demand estimation is crucial to assess the energy performance in buildings and represents the first step to make any decision for enhancing their efficiency. The proposed approach has the advantage of learning a model from data about previous certificates that is then applied to new flats. The methodology can concretely help domain experts to evaluate the possible improvements of energy efficiency of flats. To this purpose, data driven models are useful for quickly estimating the expected building energy demand and in setting credible targets for improving performance [56]. In general, designers and authority planners should exploit such tools capable to suggest them where put their effort, among large stocks of buildings, and which could be the most convenient retrofitting strategies. In this way it is possible to plan future financial investment policies that leverage on specific building features and help devising more targeted actions to improve energy efficiency for different segments of buildings. Moreover the proposed methodological process allows to extract, by means of interpretable models (i.e., decision trees), useful and understandable knowledge regarding the expected energy performance of buildings according to few physical driving variables . Such benchmarks should be the reference for the building owners to improve the energy performance when it is poor and for technicians to identify the optimal cost-effective energy saving opportunities.

Author Contributions: The research presented in this paper was a collaborative effort made by all the authors. All the authors contributed to the literature review, methodology, implementation and experimental analyses, as well as to the writing and reviewing of the paper.

Funding: This research received no external funding.

Acknowledgments: The authors express their gratitude to Giovanni Nuvoli (Settore Sviluppo Energetico Sostenibile Regione Piemonte) and to CSI Piemonte.

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

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
BREEAM	Building Research Establishment Environmental Assessment Method
CASBEE	Comprehensive Assessment System for Building Environmental Efficiency
DD	Degree Days
DGNB	Deutche Gesellschaft fur Nachhaltiges Bauen of Germany
EPBD	Energy Performance of Buildings Directive
EPC	Energy Performance Certificate
EPI	Energy Performance Indicator
EUI	Energy Use Intensity
GA	Genetic Algorithm
GBL	Green Building Label of China
HK-BEAM	Hong Kong-Building Environmental Assessment Method
IEA	International Energy Agency
IQR	Interquartile Range
LEED	Leadership in Energy and Environmental Design
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MRA	Multiple Regression Analysis
MSE	Mean Square Error
NABERS	National Australian Built Environment Rating System
PED	Primary Energy Demand
PED_h	Primary Energy Demand for space heating
PED_w	Primary Energy Demand for hot water
REPT	Reduced Error Pruning Tree
RF	Random Forest
RMSE	Root Mean Square Error
SVM	Support Vector Machines

References

1. IEA. *International Energy Agency, Energy Efficiency Indicators Highlights*; OECD/IEA: Paris, France, 2016.
2. European Parliament CotEU. Directive 2010/31/EU of 19 May 2010 on the Energy Performance of Buildings (Recast). *Off. J. Eur. Union* **2010**, *53*, L153/13.
3. Andaloro, A.P.; Salomone, R.; Ioppolo, G.; Andaloro, L. Energy certification of buildings: A comparative analysis of progress towards implementation in European countries. *Energy Policy* **2010**, *38*, 5840–5866. [CrossRef]
4. Li, Y.; Chen, X.; Wang, X.; Xu, Y.; Chen, P.H. A review of studies on green building assessment methods by comparative analysis. *Energy Build.* **2017**, *146*, 152–159. [CrossRef]
5. Darko, A.; Chan, A.P. Critical analysis of green building research trend in construction journals. *Habitat Int.* **2016**, *57*, 53–63. [CrossRef]
6. Wang, S.; Yan, C.; Xiao, F. Quantitative energy performance assessment methods for existing buildings. *Energy Build.* **2012**, *55*, 873–888. [CrossRef]
7. Koo, C.; Park, S.; Hong, T.; Park, H.S. An estimation model for the heating and cooling demand of a residential building with a different envelope design using the finite element method. *Appl. Energy* **2014**, *115*, 205–215. [CrossRef]
8. Fan, Y.; Xia, X. An optimization model for building envelope retrofit considering energy performance certificate. In Proceedings of the 2017 36th Chinese Control Conference (CCC), Dalian, China, 26–28 July 2017; pp. 2750–2755.
9. Prieler, M.; Leeb, M.; Reiter, T. Characteristics of a database for energy performance certificates. *Energy Procedia* **2017**, *132*, 1000–1005. [CrossRef]

10. Yu, Z.; Haghighat, F.; Fung, B.C.; Yoshino, H. A decision tree method for building energy demand modeling. *Energy Build.* **2010**, *42*, 1637–1646. [CrossRef]
11. Pasichnyi, O.; Wallin, J.; Levihn, F.; Shahrokni, H.; Kordas, O. Energy performance certificates—New opportunities for data-enabled urban energy policy instruments? *Energy Policy* **2019**, *127*, 486–499. [CrossRef]
12. ISO 13790. *Thermal Performance of Buildings, Calculation of Energy Use for Space Heating*; International Organization for Standardization: Geneva, Switzerland, 2008.
13. UNI TS 11300-1. *Prestazioni energetiche degli edifici—Parte 1: Determinazione del fabbisogno di energia termica dell'edificio per la climatizzazione estiva ed invernale*; Standard, UNI—Ente Nazionale Italiano di Unificazione: Italy, 2014.
14. UNI TS 11300-2. *Prestazioni energetiche degli edifici—Parte 2: Determinazione del fabbisogno di energia primaria e dei rendimenti per la climatizzazione invernale, per la produzione di acqua calda sanitaria, per la ventilazione e per l'illuminazione in edifici non residenziali*; Standard, UNI—Ente Nazionale Italiano di Unificazione: Italy, 2014.
15. Di Corso, E.; Cerquitelli, T.; Piscitelli, M.S.; Capozzoli, A. Exploring energy certificates of buildings through unsupervised data mining techniques. In Proceedings of the 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Exeter, UK, 21–23 June 2017; pp. 991–998.
16. Capozzoli, A.; Serale, G.; Piscitelli, M.S.; Grassi, D. Data mining for energy analysis of a large data set of flats. *Proc. Inst. Civ. Eng. Eng. Sustain.* **2017**, *170*, 3–18. [CrossRef]
17. Cerquitelli, T.; Corso, E.D.; Proto, S.; Capozzoli, A.; Bellotti, F.; Cassese, M.G.; Baralis, E.; Mellia, M.; Casagrande, S.; Tamburini, M. Exploring energy performance certificates through visualization. In Proceedings of the Workshops of the EDBT/ICDT 2019 Joint Conference, EDBT/ICDT 2019, Lisbon, Portugal, 26–29 March 2019.
18. Fabbri, K. Planning a Regional Energy System in Association with the Creation of Energy Performance Certificates (EPCs), Statistical Analysis and Energy Efficiency Measures: An Italian Case Study. *Buildings* **2013**, *3*, 545–569. [CrossRef]
19. Pérez-Lombard, L.; Ortiz, J.; Gonzàlez, R.; Maestre, I.R. A review of benchmarking, rating and labelling concepts within the framework of building energy certification schemes. *Energy Build.* **2009**, *41*, 272–278. [CrossRef]
20. Nikolaou, T.; Kolokotsa, D.; Stavrakakis, G.; Apostolou, A.; Munteanu, C. Review and State of the Art on Methodologies of Buildings' Energy-Efficiency Classification. In *Managing Indoor Environments and Energy in Buildings with Integrated Intelligent Systems*; Springer International Publishing: Cham, Switzerland, 2015; pp. 13–31.
21. Lu, X.; Lu, T.; Kibert, C.J.; Viljanen, M. A novel dynamic modeling approach for predicting building energy performance. *Appl. Energy* **2014**, *114*, 91–103. [CrossRef]
22. Tronchin, L.; Fabbri, K. Energy performance building evaluation in Mediterranean countries: Comparison between software simulations and operating rating simulation. *Energy Build.* **2008**, *40*, 1176–1187. [CrossRef]
23. Patiño-Cambeiro, F.; Bastos, G.; Armesto, J.; Patiño-Barbeito, F. Multidisciplinary Energy Assessment of Tertiary Buildings: Automated Geomatic Inspection, Building Information Modeling Reconstruction and Building Performance Simulation. *Energies* **2017**, *10*, 1032. [CrossRef]
24. Melo, A.; Cóstola, D.; Lamberts, R.; Hensen, J. Development of surrogate models using artificial neural network for building shell energy labelling. *Energy Policy* **2014**, *69*, 457–466. [CrossRef]
25. Dall'O', G.; Sarto, L.; Sanna, N.; Tonetti, V.; Ventura, M. On the use of an energy certification database to create indicators for energy planning purposes: Application in northern Italy. *Energy Policy* **2015**, *85*, 207–217. [CrossRef]
26. Chung, W.; Hui, Y.; Lam, Y.M. Benchmarking the energy efficiency of commercial buildings. *Appl. Energy* **2006**, *83*, 1–14. [CrossRef]
27. Hong, T.; Koo, C.; Kim, D.; Lee, M.; Kim, J. An estimation methodology for the dynamic operational rating of a new residential building using the advanced case-based reasoning and stochastic approaches. *Appl. Energy* **2015**, *150*, 308–322. [CrossRef]
28. De Ruggiero, M.; Forestiero, G.; Manganelli, B.; Salvo, F. Buildings Energy Performance in a Market Comparison Approach. *Buildings* **2017**, *7*, 16. [CrossRef]

29. Gao, X.; Malkawi, A. A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm. *Energy Build.* **2014**, *84*, 607–616. [CrossRef]
30. Lara, R.A.; Pernigotto, G.; Cappelletti, F.; Romagnoni, P.; Gasparella, A. Energy audit of schools by means of cluster analysis. *Energy Build.* **2015**, *95*, 160–171. [CrossRef]
31. Collins, M.; Curtis, J. Bunching of residential building energy performance certificates at threshold values. *Appl. Energy* **2018**, *211*, 662–676. [CrossRef]
32. Lin, M.; Afshari, A.; Azar, E. A data-driven analysis of building energy use with emphasis on operation and maintenance: A case study from the UAE. *J. Clean. Prod.* **2018**, *192*, 169–178. [CrossRef]
33. van den Brom, P.; Meijer, A.; Visscher, H. Performance gaps in energy consumption: household groups and building characteristics. *Build. Res. Inf.* **2018**, *46*, 54–70. [CrossRef]
34. Droutsa, K.G.; Kontoyiannidis, S.; Dascalaki, E.G.; Balaras, C.A. Mapping the energy performance of hellenic residential buildings from EPC (energy performance certificate) data. *Energy* **2016**, *98*, 284–295. [CrossRef]
35. Khayatian, F.; Sarto, L.; Dall'O', G. Application of neural networks for evaluating energy performance certificates of residential buildings. *Energy Build.* **2016**, *125*, 45–54. [CrossRef]
36. Park, H.S.; Lee, M.; Kang, H.; Hong, T.; Jeong, J. Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques. *Appl. Energy* **2016**, *173*, 225–237. [CrossRef]
37. Tso, G.K.; Yau, K.K. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy* **2007**, *32*, 1761–1768. [CrossRef]
38. Magalhães, S.M.; Leal, V.M.; Horta, I.M. Modelling the relationship between heating energy use and indoor temperatures in residential buildings through Artificial Neural Networks considering occupant behavior. *Energy Build.* **2017**, *151*, 332–343. [CrossRef]
39. Acquaviva, A.; Apiletti, D.; Attanasio, A.; Baralis, E.; Bottaccioli, L.; Castagnetti, F.B.; Cerquitelli, T.; Chiusano, S.; Macii, E.; Martellacci, D.; et al. Energy Signature Analysis: Knowledge at Your Fingertips. In Proceedings of the 2015 IEEE International Congress on Big Data, New York, NY, USA, 27 June–2 July 2015; pp. 543–550.
40. Fabbri, K.; Tronchin, L.; Tarabusi, V. Real Estate market, energy rating and cost. Reflections about an Italian case study. *Procedia Eng.* **2011**, *21*, 303–310. [CrossRef]
41. Hjortling, C.; Björk, F.; Berg, M.; af Klintberg, T. Energy mapping of existing building stock in Sweden—Analysis of data from Energy Performance Certificates. *Energy Build.* **2017**, *153*, 341–355. [CrossRef]
42. Xiao, H.; Wei, Q.; Jiang, Y. The reality and statistical distribution of energy consumption in office buildings in China. *Energy Build.* **2012**, *50*, 259–265. [CrossRef]
43. Dascalaki, E.G.; Kontoyiannidis, S.; Balaras, C.A.; Droutsa, K.G. Energy certification of Hellenic buildings: First findings. *Energy Build.* **2013**, *65*, 429–437. [CrossRef]
44. Gangolells, M.; Casals, M.; Forcada, N.; Macarulla, M.; Cuerva, E. Energy mapping of existing building stock in Spain. *J. Clean. Prod.* **2016**, *112*, 3895–3904. [CrossRef]
45. MISE. *Decreto Ministeriale 26/6/2009—Ministero dello Sviluppo Economico. Linee guida nazionali per la certificazione energetica degli edifici*; MISE-Ministero dello Sviluppo Economico: Roma, Italy, 2009.
46. Hastie, T.; Tibshirani, R.; Friedman, J. Boosting and Additive Trees. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009; pp. 337–387.
47. Hofmann, M.; Klinkenberg, R. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2013.
48. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
49. RapidMiner. RapidMiner Operator Reference Manual. Available online: https://docs.rapidminer.com/latest/studio/operators/ (accessed on 1 April 2019).
50. Vapnik, V.N.; Vapnik, V. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998; Volume 1.
51. Ben-Hur, A.; Weston, J. A User's Guide to Support Vector Machines. In *Data Mining Techniques for the Life Sciences*; Carugo, O., Eisenhaber, F., Eds.; Humana Press: Totowa, NJ, USA, 2010; pp. 223–239.
52. Thaseen, S.; Kumar, C.A. An analysis of supervised tree based classifiers for intrusion detection system. In Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, Salem, India, 21–22 February 2013; pp. 294–299.
53. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

54. Pang-Ning T.; Steinbach M.; Kumar V. *Introduction to Data Mining*; Addison-Wesley: Boston, MA, USA, 2006.
55. Rafiq, M.; Bugmann, G.; Easterbrook, D. Neural network design for engineering applications. *Comput. Struct.* **2001**, *79*, 1541–1552. [CrossRef]
56. Capozzoli, A.; Grassi, D.; Causone, F. Estimation models of heating energy consumption in schools for local authorities planning. *Energy Build.* **2015**, *105*, 302–313. [CrossRef]