Bridging the gap between complexity and interpretability of a data analytics-based process for benchmarking energy performance of buildings

Antonio Galli[a], *Marco Savino Piscitelli*[b1], *Vincenzo Moscato*[a], *Alfonso Capozzoli*[b]

[a] *Department of Electrical Engineering and Information Technology (DIETI), University of Naples "Federico II", Via Claudio 21, Naples, Italy*

[b] *Department of Energy (DENERG), TEBE Research Group, BAEDA Lab, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Turin, Italy*

[1]Corresponding author

E-mail address: antonio.galli@unina.it (A. Galli), marco.piscitelli@polito.it (M. S. Piscitepllli), vmoscato@unina (V. Moscato), alfonso.capozzoli@polito.it (A. Capozzoli)

Phone number: +39 081 768.3835 (A.Galli and V.Moscato),+39 011 090.4554 (M. S. Piscitelli), +39 011 090.4413 (A. Capozzoli)

Bridging the gap between complexity and interpretability of a data analytics-based process for benchmarking energy performance of buildings

Abstract

Artificial intelligence (AI) is fast becoming a general purpose technology with outstanding impacts in industries worldwide, thus supporting the Industry 4.0 revolution. In particular, the energy sector is one of those that has taken more advantages from the implementation of AI approaches, especially Machine Learning models, for several applications, including energy performance benchmarking of buildings. However, the black-box approach could lead to a lack of result interpretability thus preventing the effective application of AI in some real-world scenarios. For this reason, eXplainable Artificial Intelligence (XAI) tools can be effectively embedded within an AI-based Energy Analytics methodology in order to enhance the explainability of the model results. In this paper, we propose an explainable AI-based benchmarking framework for estimating the membership to specific energy performance classes of a large set of Energy Performance Certificates (EPCs) of flats. The classification is obtained by leveraging different black-box classifiers characterized by high accuracy, yet their inference mechanism is not human-readable. Therefore, a generalizable XAI methodology, based on the combination of a local explainer together with a clustering algorithm, is employed to explain the model results and causal effects between the predictors and target variable to better understand the model behavior, and the motivations behind correct and wrong performed classifications. The paper provides a general methodological approach capable to exploit a limited number of instances to extract, explain and interpret inference mechanisms learnt by the model that are useful for the end-user. The framework was tested on about 100,000 EPCs of flats located in Italy.

Keywords: Building energy benchmarking, Energy performance certificates, Classification algorithms, Clustering analysis, Explainable artificial intelligence.

1. Introduction

The building sector is recognised as one of the largest primary energy consumers worldwide. According to the International Energy Agency (IEA) among the EU member countries, buildings are responsible for about the 21% of total final energy consumption (Millar et al., 2016). Specifically, more than the 50% of this energy amount is used by heating and cooling systems installed in residential buildings (Millar et al., 2016). As a consequence, the building sector is currently one of the most strategic targets for decreasing overall energy demand, improving energy efficiency to achieve demanding decarbonisation objectives.

In this context, energy benchmarking systems play a key role in the evaluation of the energy performance of buildings supporting different stakeholders (public and private) in the process of energy management and planning for achieving energy saving objectives. Cities around the world began benchmarking their building stock after realizing the potential of energy benchmarking systems and recorded an energy saving up to 8% in a reference period of 3-4 years from their implementation (Arjunan et al., 2022; Frick et al., 2017).

From the technical point of view, the main goal of a benchmarking system is to evaluate, in a systematic way, the divergence between the energy performance of a building/system and a reference baseline. Four types of baselines can be considered in existing benchmarking methods: previous performance of similar buildings (i.e., external benchmarking), current/intended performance of similar buildings (i.e., external benchmarking), previous performance of the same building (i.e., internal benchmarking), and intended performance of the same building (i.e., internal benchmarking) (Li et al., 2014). The first two types of baselines are used by regulators, public authorities, or private building portfolio managers to encourage owners to improve energy efficiencies of their buildings (Chung, 2011). On the other hand, internal benchmarking techniques are exploited at single building level for energy performance tracking and continuous commissioning purpose.

According to the modeling approach considered, benchmarking systems can be further classified in calculation-based and data-driven ones (Wang et al., 2012). The calculation-based benchmarking system compares the observed energy consumption with a simulated benchmark, representing an

archetype or a theoretical energy performance (Lee et al., 2003). Simulation tools, belonging to the so-called white box methods, are by now the main instrument to assess the energy performance of buildings and to evaluate the possible scenarios for energy retrofit (Fabrizio et al., 2010; Hong et al., 2015; Mauro et al., 2015; Lee et al., 2015; Tahsildoost & Zomorodian, 2015); they also provide the most reliable results at the design stage of a building (Al-Homoud, 2001). This approach was however of limited use for large building stocks because it is time-consuming, labour intensive (Filogamo et al., 2014), and it requires detailed building information which is not always easily available at large scale (Zhang et al., 2015). On the other hand, the data-driven benchmarking process compares the observed energy consumption with a benchmark value obtained from actual energy consumption data. The most common data-driven benchmarking processes, proposed in the literature, are performed through statistical models (Lee & Lee, 2009), data analytics techniques (Petcharat et al., 2012; Capozzoli et al., 2016a; Gao & Malkawi, 2014) and simple normalization of the energy consumption with respect to floor area and/or volume as a way to compute the mean or median value (Wang et al., 2012).

With the rapid growth of stored and open data in building sector and the necessity to extract knowledge from these large data sets to improve the building performance, data-driven benchmarking systems are more and more emerging (Papadopoulos & Kontokosta, 2019; Yang et al., 2018; Roth et al., 2020). The choice of the most suitable strategy (simple normalization, statistical models or data analytics techniques) to develop a benchmarking process mainly depends on the quantity and the quality of the available information and on properties of the considered dataset.

In the last decades, instruments, such as, Energy Performance Certificates (EPCs) have emerged as a key tool for driving the definition of energy efficiency policies for the building sector. As a reference, under the Energy Performance Buildings Directive (EPBD) (2002/91/EC), EPCs have become compulsory in EU Member States. The EPBD allows 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). EPCs provide theoretical measure of building performance if they are operated in standard conditions. However, the performance gap, i.e. the difference between estimated and actual energy performance could be significant. For instance, Pasichnyi et al. (2019) stated

that for the Swedish EPCs data-set the 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 conduct comparisons between a building and its peers.

As emerged from the scientific literature, EPCs data sets represent today great sources of information and a growing number of researchers are using them for addressing different tasks in the context of building energy management (Pasichnyi et al., 2019) including advanced benchmarking analysis (Attanasio et al., 2019). The interest in energy performance assessment is increased especially to estimate how the combination of different features affects the energy needs in buildings (Arjunan et al., 2020). In fact, from the design point of view, it is crucial to determine the effect of the building features on its future energy performance in the early design phase. Similarly, for existing buildings, it could be useful to evaluate the feasibility and impact of a refurbishment plan. Regardless the scope to be pursued, estimating building energy performance in a quick and reliable way, for different combinations of building features, is essential for different actors such as building owners, designers, facility managers and public authorities (Attanasio et al., 2019; Capozzoli et al., 2015). Despite this, building professionals are typically suspicious towards the prediction results of data-driven processes because they cannot always fully interpret the model inference mechanism. In fact, what not-expert users need in practice is not only the result obtained through a single prediction, but also explanations for improving the awareness of the decision-making process. In this perspective it is becoming more and more important to develop predictive analytics tools capable of providing feedbacks about the reasons behind a certain prediction with robust indication of the supporting and conflicting evidences towards it (Fan et al., 2019b; Miller, 2019b; Arjunan et al., 2022).

Explainable artificial intelligence, also called XAI, is an emerging subject in the field of big data analytics. It aims to provide methods and tools to enhance the model usability breaking the trade-off between model complexity and model interpretability (Fan et al., 2019b). Considering the practical difficulties faced by building professionals in utilized advanced supervised learning techniques, XAI is very promising to fully exploit the potential of advanced machine learning techniques in the building application field (Sardianos et al., 2021; Miller, 2019b; Arjunan et al., 2022).

According to the aforementioned motivations, the objective of this work is twofold. Firstly, the work proposes a data-driven process capable to es-

timate, for a large set of EPCs of flats, the membership to specific energy performance classes for benchmarking purpose. The classification task is performed through different ML classifiers characterised by high accuracy but whose inference mechanism, despite in some cases is human-readable, can not be easily interpreted by the end-user. Secondly, a XAI-based explanation process is introduced to provide insight about the behaviour of classification models used to benchmark the energy performance of buildings and to understand the motivations behind correct and wrong classifications (this information can be very helpful for e.g., certification entities or technical figures). To this aim, the explanation process combines the XAI tool called LIME together with k-means clustering method for providing local but representative explanations of model predictions. The proposed methodology was then tested on an EPC data-set related to about 100,000 flats located in Piedmont (north-western region in Italy).

In the light of the objective of this paper, the following Section 2 reports and discuss the literature concerning the implementation of XAI-based processes in different fields of research including the one of energy and buildings. The main contributions to the literature, an the novelty introduced with this study, are presented and discussed in 3.

2. Related works

Over the last two decades,the robust coupling of artificial intelligence (AI) and energy domain knowledge proved to be effective in achieving relevant energy saving in buildings by exploiting a variety of predictive-based energy management solutions, such as energy consumption forecasting (Fan et al., 2019a; Sun et al., 2020; Runge & Zmeureanu, 2019), anomaly/fault detection and diagnosis in buildings and energy systems (Himeur et al., 2021; Piscitelli et al., 2020), advanced energy benchmarking (Geraldi & Ghisi, 2022; Li et al., 2020), load profiling (Eskandarnia et al., 2022; Liu et al., 2021). Predictive analytics is de facto considered a cross-sectional application of AI for enhancing energy management in buildings (Zhao et al., 2020), and until now its use has been associated to the need of achieving the highest accuracy as possible of predictions at the basis of decision-making process.

However understanding *why* a certain prediction is provided by a black-box model is becoming more and more an essential feature of predictive analytics in several modern contexts, especially when the decisions of an AI system are required to be transparent and fair (e.g, for certification aims).

Generally speaking, such task is the main goal of *eXplainable Artificial Intelligence* (XAI) (Gunning, 2017), which offers new opportunities for successfully embedding AI-based solutions in industrial applications where explanations of the data-driven AI models is often a mandatory requirement.

In the last decade, XAI has become for AI researchers an emerging and very challenging topic whose meaning and usefulness can be summarized through several key aspects, as reported by the first significant published studies (Ribeiro et al., 2016; Biran & Cotton, 2017). Certainly, the two most relevant concepts concern the ability of an AI system to explain its decisions in intelligible terms to humans (Došilović et al., 2018; Guidotti et al., 2018) (i.e., *Explainability*) and, the ability to identify the set of characteristics that mostly contribute to making a decision (Adadi & Berrada, 2018) (i.e., *Interpretability*). Therefore, XAI supports the definition of more explainable models, maintaining a high level of performances, allowing human users to understand and trust AI-based systems.

During the last years, different XAI methods and strategies have been proposed. Usually, they can be classified, according to the granularity of the related analysis, into *local* (understand a single prediction) and *global* (understand the model behaviour) approaches.

Local XAI methods aim to explain a black-box model outcome on the basis of local information around the prediction. For instance, Baehrens et al. (2010) have proposed to measure local gradients to exactly identify in which ways changing the input affects the prediction. Similarly, Robnik-Šikonja & Kononenko (2003) presented a feature importance method, which computes the differences between a prediction and the obtained solution. Finally, the model-agnostic (LIME) method proposed by Ribeiro et al. (2016) is based on an algorithm that faithfully explains the predictions of any classifier, by approximating it locally with a fully interpretable model.

The techniques above summarized focus on local explanations to achieve an overall explanation of a model. On the other hand, other techniques explicitly try to build global explanations. The most popular methods of this latter typology rely on features importance to explain tree-based models: the global Mean Decrease in Impurity (MDI) approach (Breiman, 2001) – which exploits splits' number of samples – and the Mean Decrease in Accuracy (MDA) technique (Louppe, 2014) – which computes a model mean increase error on the basis of a random permutation of the features.

Nowadays, many researchers are introducing further XAI methodologies as part of the development loop of machine learning models, in applications

related to several domains. For example, Pereira et al. (2018) used a Grad-CAM and Guided BackPropagation (GBP) (Springenberg et al., 2014) to analyse the clinical coherence of the features learned by a CNN for automated grading of brain tumours in magnetic resonance imaging. More recently, Chen & Lee (2020) exploited a Gradient Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) to generate the explainability map of a model, showing that it focused its attention in high-dimensional bands excitepd by structure resonance. Galli et al. (2020) proposed a framework to predict hard disk drives health status including an effective explanations of the model decisions leveraging different XAI tools. Chakraborty et al. (2021) compared the predictive capabilities of three interpretable ML models with three non-interpretable ML models using measured climate data and two different XAI tools.

Recently, also in the energy domain the concept of XAI is being introduced. Fan et al. (2019b) defined a comprehensive methodology to explain and evaluate building energy performance models. Whereas, Arjunan et al. (2020) introduced a methodology that enhances the existing building benchmarking process of Energy Star by increasing accuracy and providing additional model output processing to help explaining why a building is achieving a particular energy performance score. Akhlaghi et al. (2020) developed an explainable and interpretable Deep Neural Network (DNN) model for a Guideless Irregular Dew Point Cooler (GIDPC). The SHapley Additive exPlanations (SHAP) method was used to assess and interpret the contribution of the operating conditions on performance parameters of the system. Also Kuzlu et al. (2020) employed a XAI-based process for improving the interpretability of a prediction model. In particular, the study focused on the forecasting of solar PV system generation and on the use of XAI tools, such as LIME, SHAP, and ELI5, for model explanations. A further promising application of XAI in the energy and buildings field concerns with fault detection and diagnosis (FDD). In particular, XAI offers the opportunity to explain which are the boundary conditions related to the detection of a fault/anomaly during system operation and most of all provides a readable interpretation about its diagnosis. As a reference, an interesting application of XAI was proposed by Madhikermi et al. (2019) for enhancing FDD analysis based on machine-learning algorithms (i.e., support vector machine, artificial neural network) conducted on building Air Handling Unit (AHU).

3. Novelty and contribution of the work

The work presented in this paper aims to introduce a novel approach in explanation analysis leveraging local XAI tools, such as LIME, for providing insights about the behaviour of classification models used for benchmarking the energy performance of buildings.

The approach combines different advanced data analytics techniques with the aim of maintaining the output of an external building energy benchmarking process human-readable and interpretable while providing accurate and reliable results. This aspect is extremely valuable for such kind of benchmarking systems because it is usually employed by regulators, public authorities, or managers involved in the decision-making process of large building portfolio energy management. For this reason the proposed approach was tested on an EPCs dataset related to the energy performance evaluated for about 100,000 flats located in Piedmont (north-western region in Italy).

Despite the spread of XAI techniques in several domains, to the best of authors' knowledge, the proposed approach represents the first attempt to investigate the problem of explainability in the Energy Analytics domain for the automatic estimation of building energy performances using an EPCs dataset. In particular, different ML models were firstly developed in order to solve the classification task under analysis (estimation for a new instance of its membership to an energy performance class). Successively, a XAI-based process was used to probe the rationale behind model decisions in order to understand the motivations behind right and wrong classifications.

In this context the main innovative aspects introduced by the present paper can be summarised as follows:

- The proposed framework makes it possible to employ the best classifier for energy benchmarking (in terms of achieved accuracy) regardless to its level of interpretability. From the energy point of view, the interpretability of the data-driven model used for benchmarking building energy performance is often considered a constraint in the selection of the prediction model to be used and can have repercussion on the final achievable accuracy (Capozzoli et al., 2016b,a; Attanasio et al., 2019). Following the proposed approach the end-user has the possibility to easily probe the rationale behind model decisions following an agnostic approach and then to select the model that can better extract the main patterns from data, achieving the best accuracy.

- In this study a detailed analysis was performed for better understanding the behaviour of the model in handling classifications in border areas across adjacent energy performance classes. The analysed dataset is particularly dense and includes about 100,000 EPCs of flats. As a consequence, each energy performance class can not be considered well separated from the adjacent ones. In this context the main targets of the explanation analysis are the instances misclassified due to border effects in order to understand the model behavior and assess the trustworthiness of its predictions in such particular cases.

- The explanation analysis has been conceived both to support the end-user during the deployment phase of the benchmark model (i.e., explanation at single prediction level) and to guide the analyst in extracting the macro-behaviors of the classification models under particular conditions (i.e., misclassification of border objects). The latter objective has been pursued by coupling the local explanation algorithm, i.e., LIME with a k-means clustering analysis, in order to firstly recognise the most significant groups of similar instances in the dataset and then explain predictions with reference to prototype objects (i.e., cluster centroids) that have been intended as representative of groups of instances. In this way, the analyst is provided with a set of reference explanations to assess some key feature combinations that could lead to more certain/uncertain predictions.

The rest of the paper is organized as follows. Section 4 provides an overview and a brief theoretical description of the data analytics methods used for conducting the analysis. Section 5 presents and describes the case study considered for the analysis. Section 6 introduces the methodological framework behind the analysis performed. Eventually, Sections 7 and 8 present and discuss the results obtained while in Section 9 the concluding remarks and future research perspectives are reported.

4. Materials and Methods

In this section, the data analytics methods employed in this work are briefly described. The method descriptions are not intended to be exhaustive, but they are aimed to underline the main model features according to the objectives of this study. In particular, the classification algorithms

used for developing the energy benchmarking model based on EPCs are described. Successively, a brief introduction to k-means clustering technique is provided. Eventually, the main theoretical principles of the LIME explanation algorithm are reported.

4.1. Classification Algorithms

As well known, classification is related to a predictive modeling problem where a class label has to be predicted starting from labelled input data. There exists a plethora of classification algorithms that can be conveniently used depending on the dataset features. Below, the five algorithms exploited in the present study (i.e., Decision tree, Random Forest, Extremely Randomized Tree, Bagging classifier, MultiLayer Perceptron) were introduced and described.

A *Decision Tree* (DT) (Quinlan, 1986) is a supervised learning algorithm that fits well with many kinds of classification problems. Given a set of labelled data, a decision tree produces a sequence of IF-THEN rules that can be used to classify the data. It works like a flow chart, separating data points into two similar categories at a time from the "tree trunk" to "branches", to "leaves", where the categories become more finitely similar. DT requires little data preparation, and can handle both numerical and categorical data. However, it can create complex trees that do not generalize well, and can be unstable because small variations in the data might result in a completely different tree being generated.

The *Random Forest* (RF) (Breiman, 2001) algorithm is an expansion of the DT concept. The term "forest" is referred to an ensemble of DTs, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. Specifically, it develops a number of DTs on various sub-samples of the dataset and uses average to improve the predictive accuracy of the model and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

Similar to RF classifier, in this study were also used the *Extra Trees* (ET) (Geurts et al., 2006) classifiers — also known as *Extremely Randomized Trees*. The two ensembles have a lot in common. Both of them are composed by a large number of DTs, where the final decision is obtained taking into account the prediction of every tree. The main differences are the following: i) RF uses replicas, it subsamples the input data with replacement, whereas ET

uses the whole original sample; ii) RF chooses local optimum splits while ET chooses it randomly.

Bagging Classifier (BC) (Breiman, 1996) is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the misclassification error of a black-box estimator, by introducing randomization into its development procedure and then making an ensemble out of it. The total expected error of a classifier is made up of the sum of two components, the bias and the variance. More in details, the bias for a learning rate problem is the error rate for a particular learning algorithm and measures how well the learning method matches the problem. Since the used training set is finite and not fully representative of the population of instances, a second source of error is inevitably introduced. The variance is the expected value of this component of error, over all possible training sets and test sets. Combining multiple classifiers generally decreases the total expected error by reducing the variance component: the more classifiers that are included, the greater the reduction in variance.

Eventually, a *MultiLayer Perceptron (MLP)* was implemented. In particular, an MLP is a class of feedforward artificial neural network (ANN) and consists of at least three layers of nodes: i)input layer; ii) hidden layer and iii) output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training.

As demonstrated by the recent literature (Amasyali & El-Gohary, 2018; Miller, 2019a), the presented classification approaches are some of the most spread ones in the energy analytics field, especially for building energy performance assessment and load forecasting.

4.2. Clustering

Clustering consists in grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters (Tan et al., 2013). In this study, the similarity between objects was based on a measure of the Euclidian distance (Eq. 1), as follows:

$$d(x, y) = \sqrt[2]{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

where, x and y are two vectors of length n representing the samples.

K-means is a partitive clustering algorithm (Tan et al., 2013) that consists in grouping data objects into non-overlapping subsets (i.e., clusters) such that each data object can be included only in one sub-set. K-means is used for grouping data objects in a pre-determined number of K clusters which are represented by a prototype object called centroid (i.e., mean of the points in the n-dimensional space). The first step of K-means consists in the setting of the number K of desired clusters to which corresponds a prototype object (centroid) randomly located in the n-dimensional space (Tan et al., 2013). Each object in the dataset is then assigned to the closest centroid, and each group of objects assigned to the same centroid represents a cluster. The centroid of each cluster is then recalculated as the average of all the objects assigned to the cluster. This process is repeated until the data objects do not change cluster anymore, and the centroids do not change position.

4.3. Local agnostic explaination analysis with LIME

One of main features of LIME is its modeling-agnostic nature, that makes the XAI tool applicable to any ML model. More in detail, LIME provides the local interpretability of a model: each instance is fed into the model providing both a prediction and a local sensitivity analysis with the aim of highlighting how sensitive the outcome is to each input feature.

In other words, the algorithm infers the behavior of the model by perturbing the input data and analyzing how the predictions change accordingly. In practise, the output of LIME is a list of explanations reflecting the contribution of each feature to the outcome of a given instance. As a consequence, LIME enables the local interpretability of a prediction, allowing to determine how the change of a feature will impact the model output.

The above discussed process, related to the explanation produced by LIME can be formalized as follows. Considering a local point x the relative explanation is obtained with the following generic formula (Eq. 2):

$$\xi(x) = \underset{g \in G}{argmin} \;\; L(f, g, \pi_x) + \Omega(g) \tag{2}$$

where G is a class of potential interpretable models, $\Omega(g)$ expresses a measure of the complexity of the interpretability of $g \in G$. The model being explained be denoted *f*, so *f(x)* is the probability that x belongs to a certain class. Furthermore, $\pi_x(z)$ is defined as proximity measure between an instance z to x, so as to define locality around x.

Eventually, $L(f, g, \pi_x)$ is a measure of how unfaithful g is in approximating f in the locality defined by π_x. In order to ensure both interpretability and local fidelity, we must minimize $L(f, g, \pi_x)$ while having $\Omega(g)$ be low enough to be interpretable by the human user.

5. Case study

The analysed case study pertains to EPCs issued for about 100,000 flats located in Piedmont region (Italy). The EPCs include several features that impact on the building energy performance, as well as the parameters employed to determine its energy needs. Based on the analyses on EPCs previously carried out in Capozzoli et al. (2016b) and Di Corso et al. (2017), a proper selection of important and easy-to-collect variables has been carried out. The following four main types of input variables to the benchmarking model, were identified: (i) *Geometry*, (ii) *Envelope*, (iii) *Time* and (iv) *System*.

The variables in the category *Geometry* include different geometric features of the flat, which impact on its energy need and performance. Several variables belong to this category such as the average ceiling height, the heat transfer surface and the gross heated volume of the flat.

The variables in the category *Envelope* are representative of the main physical properties of the opaque and transparent envelope of the flat (e.g., the thermal transmittance values of the opaque and transparent building envelope).

Moreover, in the category *Time* are included time variables such as the construction year of the building (in which is located the flat considered).

Lastly, the variables related to the heating system belong to the category labelled as *System* (e.g., the average overall efficiency of the system for space heating). The variable *average overall efficiency of the heating system* is calculated according to the standard efficiency values for each subsystem (i.e., generation, distribution, control, emission) reported into the part 2 of UNI/TS-11300 (2008).

Among all the variables that can be extracted from an EPC, the Primary Energy Demand for space heating PED_h has been selected as the target variable of the benchmarking analysis. PED_h (expressed in kWh/m^2y) is an energy-related variable defined for benchmarking purposes (it contributes to assign an energy class label to the flat) and consists in an estimation of the energy demand of a flat under standard use conditions. The PED_h value

pertains to the energy demand referred to a standard period of a heating season and it is normalized by the flat floor area. PED_h is part of the estimated overall Primary Energy Demand of flats (PED) which also includes the Primary Energy Demand for domestic hot water (PED_w). More specifically, the heating energy demand is assessed performing an energy balance of the flat. The modeling of the building geometry considers real shapes and self or over shading of other buildings/external obstructions. The calculation procedure considers a quasi steady-state approach based on the monthly balance of heat losses (due to transmission and ventilation) and heat gains (considering both solar and internal gains) that are evaluated in monthly average conditions. In particular the standard monthly outdoor climatic conditions (i.e., temperature and solar radiation) referred to a location on the national territory are reported in the national technical regulation UNI 10349-1. Specifically, the monthly outdoor climatic conditions, reported in the part 1 of UNI-10349 (2016), are evaluated according to the standard ISO-15927 (2003) which prescribes the use of at least 10 years of measured meteorological data for the calculation. The estimation of the transmission heat losses is performed considering actual stratigraphies and thermal properties of opaque and transparent envelopes and as well as the thermal bridging effect. In standard rating conditions, parametric values related to floor area or heated net volume, are used for defining the ventilation rates and internal heat gains. The dynamic effects and their influence on the net heating energy demand are modeled by introducing the dynamic parameters such as utilization factors and adjustments of the set-point temperature related to intermittent heating/cooling or set-back. These dynamic parameters are related to the building thermal inertia, the ratio between heat gains and heat losses and the occupancy/system operation schedules. From the system side, the annual PED for space heating depends on different efficiencies considering the thermal losses in the various heating sub-systems (emission, control, distribution, generation). For the heating season, the average system efficiency is calculated as the ratio between the net building energy need and the PED for heating. Furthermore, the PED also takes into account the electrical energy demand of auxiliary systems.

In order to remove the climatic effect and make flats comparable, PED_h is recalculated according to a reference standard climatic condition. More specifically, all the EPCs issued in Piedmont region provide an estimation of the PED_h for the standard climatic conditions of the actual city (in which the building is located), and for the city of Turin (i.e., Province capital). As

a consequence, the PED_h values considered in this study assume all flats as located in the same city considering then the same standard monthly outdoor climatic conditions (i.e., temperature and solar radiation). In this way, comparisons among flats are consistent. Nevertheless, if the performance rating of a flat is required to be performed for a city different from Turin, a data scaling process based on the use of standard Degree Days (DD) represents a robust approach. In particular, the scaling of the estimated PED_h can be obtained by multiplying it for the ratio between the standard DD value referred to the actual location of the flat and the ones of Turin.

6. Methodology

In this section the conceived methodology is presented and described. The proposed methodology aims to develop an energy benchmarking tool based on a classification model. Successively a XAI-based process is employed to interpret the obtained results in order to better understand the model behaviour and the motivations behind correct and wrong classifications. As described in the previous sections, develop a high performance and interpretable model is not a straightforward task and it needs to take into account several aspects. The methodology unfolds over four stages as shown in Fig. 1.

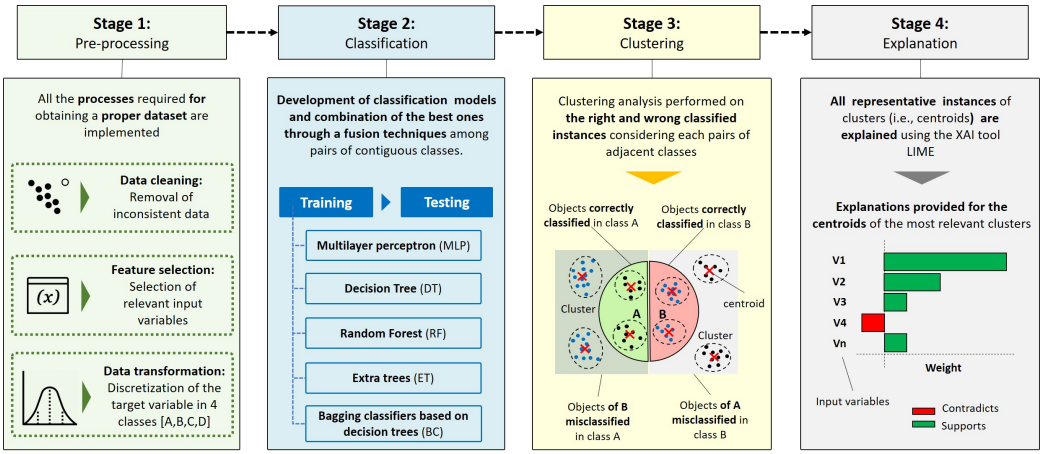


Figure 1: Methodological framework of the proposed study.

1. *Data pre-processing stage*: all the preliminary tasks necessary to provide the proper dataset to the algorithms were implemented.

2. *Classification stage*: several classification algorithms were trained and tested, with the aim of evaluating how a classification model assigns the flats to different predefined energy performance classes.
3. *Clustering stage*: a clustering analysis was performed on the correct and wrong classified instances in order to identify the most relevant predictions to be explained.
4. *Explanation stage*: all the representative instances identified by means of the clustering analysis were explained and interpreted using the XAI algorithm LIME.

In the following, each stage of the methodology is described and discussed in more detail.

6.1. Data preprocessing

The EPC dataset includes several variables of different types (numerical, categorical, textual, etc.) related to different features affecting building energy performance as well as the variables used to quantify its energy demand. Some of the available variables were not necessarily relevant for next data analysis, which means that their inclusion in the set of features could have increased the complexity of the benchmarking model. For this reason the dataset was inspected from energy domain experts, in order to remove the less relevant features for the analysis. The selection of the predictors has been driven from previous experiences collected on the same case study (Capozzoli et al., 2016b; Attanasio et al., 2019; Di Corso et al., 2017) and from the need of considering only easy-to-collect variables typically included in a EPC. All the selected features have an influence on primary energy demand from the physics point of view. It is worth to note that most of the discarded variables were poorly related with the target variable or redundant with other ones. In particular, the experiments were performed exclusively on attributes that can be categorized as geometric, thermophysical, and system-based features. The geometric and thermophysical variables are real-life variables that can be collected through surveys and inspections from energy experts before issuing an EPC. While the system-based variable (i.e., *Average global efficiency for space heating*) can be easily evaluated, with a certain degree of uncertainty, on the basis of pre-calculated values of efficiency referred to each subsystem considering the real generation system, distribution network, terminal unit, and control system installed in the building. According to this assumption, other variables that could have had a high influence on the target variable,

but are difficult to be collected or involve complex calculation procedures to be determined, were excluded.

The final set of variables, considering both inputs and output, is reported in Tab.1. For the sake of clarity, the variable aspect ratio (R) refers to the ratio of heat transfer surface area (S) to the gross heated volume (V) while the average U-values of the thermal transmittance (U_o and U_w) define the ability of the opaque and transparent envelope of the flat to transmit heat under steady-state conditions. The U-value is a measure of the quantity of heat that flows through unit area in unit time per unit difference in temperature of the environments (i.e., indoor and outdoor environment) between which the structure is located.

Category	**Name**	**Symbol**	**Unit**
	Input variables		
Geometry	Floor Area	A	m^2
	Heat transfer surface	S	m^2
	Average ceiling height	H	m
	Gross Heated Volume	V	m^3
	Aspect ratio	R	m^{-1}
Envelope	Average U-value of vertical opaque envelope	U_o	W/m^2K
	Average U-value of the windows	U_w	W/m^2K
System	Average global efficiency for space heating	η_h	$-$
	Target variable		
Energy	Normalized primary energy demand for space heating	PED_h	kWh/m^2y

Table 1: List of the input and output variables considered in the analysis.

After the feature selection, a data cleaning analysis was performed to remove statistical outliers and inconsistencies from the EPC dataset. Eventually a data transformation analysis was performed on the target variables in order to obtain a set of energy performance classes from numerical values of PED_h. Specifically, four reference classes have been considered representing respectively low energy demand flats (class A), medium energy demand flats (Class B), high energy demand flats (Class C) and very high energy demand flats (Class D). This data transformation is necessary for the construction of the classification models, which are based on a categorical response variable. The selection of threshold values between consumption classes must

be accurate to obtain reliable information from the dataset. This step was performed considering PED_h distribution and selecting as threshold values, between the classes, the 25^{th}, 50^{th} and 75^{th} percentile respectively. In this way, each energy performance class roughly includes the same number of flats avoiding then class imbalance problems that could compromise the performance of the classifiers. As a result Class A includes flats with PED_h < 91 kWh/m^2y, Class B includes flats with PED_h values between 91 and 141 kWh/m^2y, while flats in Class C have 141 $kWh/m^2y \leq PED_h < 203$ kWh/m^2y, and in Class D $PED_h \geq 203$ kWh/m^2y.

6.2. Classification analysis

The classification analysis was aimed to develop a benchmarking model capable of predicting the membership of a new flat to one of the pre-determined energy performance classes as defined above (A, B, C, D). To this purpose, the EPC dataset was grouped by contiguous class forming three binary data sets respectively $A-B$, $B-C$ and $C-D$. Each dataset has been split 80% in the training set and the remaining 20% in the testing set. Therefore, it was carried out an exploration of ML models from the simplest to the most complex one.The following classification models were trained and tested for each dataset considered in this study: Multilayer perceptron (MLP), Decision Tree (DT), Random Forest (RF), Extra Trees (ET) and Bagging Classifier based on Decision Tree (BC).

Figure 2: Graphical representation of the model selection process.

Eventually, for each of the three data sets, the algorithm which achieved the best accuracy in testing was selected as the most valuable candidate for being used as an energy performance benchmarking model. In addition, in order to uniquely select which of the three classifiers should be used for classifying an unseen flat during the deployment phase of the benchmarking tool, a model selection technique has been implemented (Fig. 2).

In particular, the model selection technique exploits class contiguity by analyzing the probabilities associated with class pairs for each model. As a reference, for a new instance the first model (i.e., binary classifier $A - B$) is used. If the probability class of A is higher than B the first model is assumed to be the most suitable for performing the prediction of the new instance, otherwise the new instance is also put through the second model. In this case, if the first and second model (i.e., binary classifier $B - C$) have both the probability of class B greater than class A and class C respectively, then the model with the highest probability of class B is chosen for performing the prediction on the new instance. Conversely, if the probability of class

C is higher than class B for the second model, then the third model is also considered. At this stage, the above described process is the same for selecting the best model among the second and third classifier (i.e., binary classifier $C - D$).

Finally, a global SHAP analysis was performed for each of the obtained models with the aim to determine the global importance of each input variable in terms of impact that it has on the model predictions. To this purpose, according to Lundberg & Lee (2017) the average of absolute shapley values per feature across the data were considered.

6.3. Clustering analysis

The next step, after the configuration of the benchmarking models, was the clustering analysis. In particular, once the trained models were obtained, the predictions on the test sets were computed. Successively, a cluster analysis was performed in order to generate high level explanations. The result of this step, consisted in the identification of clusters that were representative of correctly predicted instances and misclassified ones, considering each pair of adjacent energy performance classes. Fig. 3 shows a graphical representation of the types of instances that were clustered through a K-means algorithm. As a reference for the pair of adjacent energy performance classes A and B the four clustering analysis were performed on the following types of instances in the testing set:

- objects labelled as A and correctly predicted as A;
- objects labelled as B and correctly predicted as B;
- objects labelled as A and wrongly predicted as B;
- objects labelled as B and wrongly predicted as A;

For each type of instances, all the variables labelled as input in Tab 1 were used by a k-means algorithm for identifying ten clusters of similar objects. Successively, only the centroids of the biggest clusters were considered in the following explanation analysis. In particular, the clusters which cover at least the 90% of the entire group of instances of the same type were selected. In this way, with a local explanation of the predictions pertaining the most relevant centroids, it is possible to extract useful knowledge about the model behaviour considering a limited number of instances in a particularly dense dataset as the considered one.

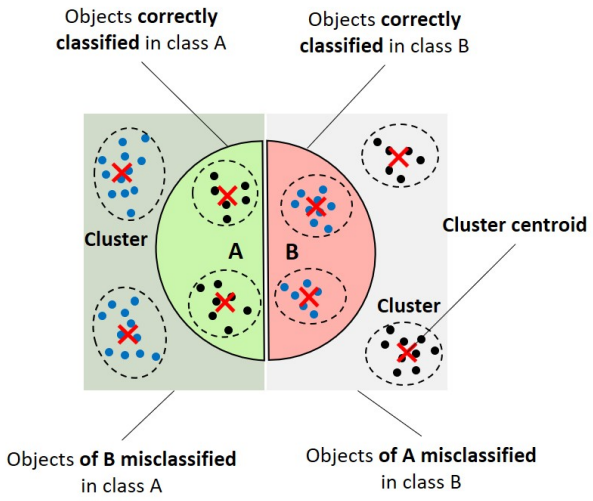


Figure 3: Graphical representation of the types of clustered instances.

6.4. Explaination analysis

Eventually, each representative instance (i.e., centroid) extracted from the previously selected clusters was explained with an XAI model-agnostic tool, that is LIME, assigning to each input variable (i.e., features of the centroid) a value of importance for a specific prediction. In particular, the main target was to explain both correct and wrong representative classifications, in order to extract, explain and interpret significant inference mechanisms learnt by the benchmarking model useful for the analyst and the end-user. In fig. 4 the typical output of the LIME tool, that explains the prediction of a binary classifier, is depicted. More in detail, the LIME output presents on the y-axis the input variable of the model ordered by decreasing importance, while on the x-axis the impact of each feature on the prediction for a given class. The use of color indicates the class towards which a feature has the highest impact.

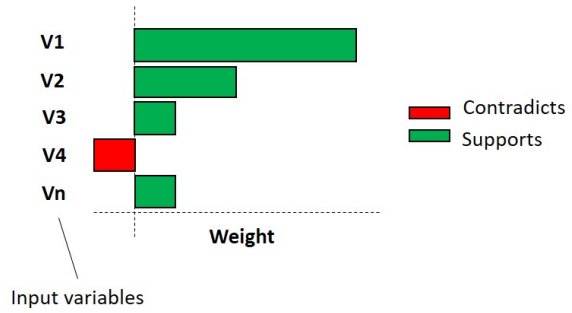


Figure 4: Graphical representation of the LIME output

7. Results

This section discusses the obtained results. The main goal is to present which are the outcomes of the proposed approach and most of all how they can be effectively used for interpreting the behaviour of the developed energy benchmarking model.

7.1. Classification Results

As previously explained in section 6, a specific pre-processing stage was considered before performing the classification task. In particular, after the data cleaning phase, the PED_h values referred to the analysed flats, were discretized and labelled with the energy performance classes A, B, C or D. The discretization was performed considering the PED_h distribution and selecting as threshold values between the classes the 25^{th}, 50^{th}, 75^{th} percentiles respectively as shown in Fig. 5.

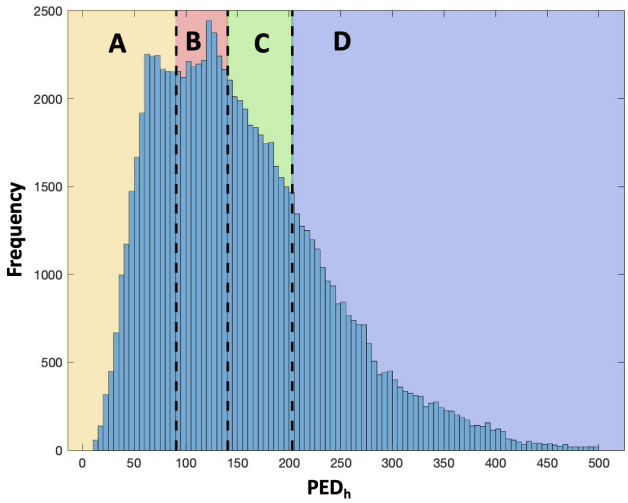


Figure 5: Identification of the energy performance classes on the PED_h distribution

Successively, the dataset was splitted 80% in the training set and 20% in the testing set and the five selected classifiers (i.e., Multilayer perceptron (MLP), Decision Tree (DT), Random Forest (RF), Extra Trees (ET) and Bagging Classifier based on Decision Tree (BC)) were developed and compared. Tab.2 shows the results obtained for each classifier in terms of accuracy, precision and recall achieved in the testing phase. In particular according to the results obtained the RF was selected as the most suitable energy benchmarking model for energy performance classes $A-B$ while the ET for the classes $B-C$ and $C-D$. Furthermore, in Tab. 3 are reported the confusion matrices related to the best models selected for each class pair.

Model	**Acc.** $A-B$	**Pre.** $A-B$	**Rec.** $A-B$	**Acc.** $B-C$	**Pre.** $B-C$	**Rec.** $B-C$	**Acc.** $C-D$	**Pre.** $C-D$	**Rec.** $C-D$
DT	74.5%	74.7%	75.0%	67.3%	67.8%	66.7%	69.3%	67.9%	**74.7%**
RF	**77.8%**	**78.6%**	77.2%	69.5%	70.4%	67.8%	70.7%	70.2%	72.9%
ET	76.9%	78.5%	74.7%	**69.6%**	**70.5%**	67.7%	**71.1%**	**71.0%**	73.0%
BC	77.2%	77.9%	76.6%	68.1%	69.1%	67.2%	68.0%	69.1%	67.2%
MLP	75.7%	71.8%	**80.5%**	67.9%	67.8%	**68.9%**	70.4%	66.6%	74.2%

Table 2: Testing accuracy (Acc.), precision (Pre.) and recall (Rec.) achieved by each developed classifier.

Successively, a model selection technique has been employed on one hand to combine the best models, and on the other hand to identify the right model to use with a new sample according to the process described in Section 6.

(a) Random Forest

	act. A	**act.** B
pred. A	3483	1035
pred. B	948	3461

(b) Extra Trees

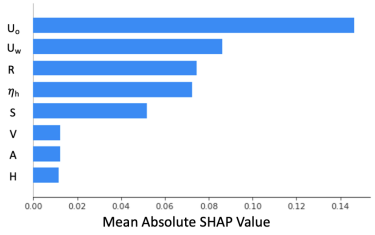
	act. B	**act.** C
pred. B	3081	1414
pred. C	1277	3155

(c) Extra Trees

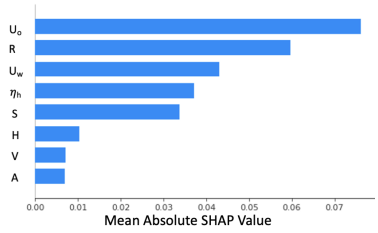
	act. C	**act.** D
pred. C	3299	1237
pred. D	1338	3068

Table 3: Confusion Matrices related to the best models, (a) Random Forest for the class pair $A-B$, (b) Extra Trees for the class pair $B-C$ and (c) Extra Trees for the class pair $C-D$.

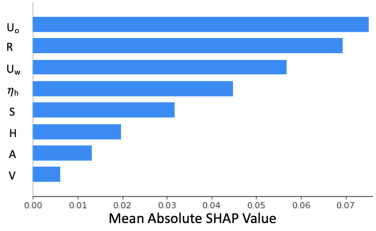
As a reference, the whole approach (i.e., based on the combination of three contiguous binary classifiers) was compared with a traditional multi-class classification approach. The proposed approach outperforms the multiclass approach by 5% in terms of overall accuracy.



((a)) Random Forest ($A-B$)



((b)) Extra Trees ($B-C$)



((c)) Extra Trees ($C-D$)

Figure 6: Global SHAP values evaluated for assessing the impact of each predictor on model output considering (a) the Random Forest for the class pair $A-B$, (b) the Extra Trees for the class pair $B-C$ and (c) the Extra Trees for the class pair $C-D$.

Eventually, in Fig. 6 the results obtained through a global SHAP analysis are reported for each of the obtained best models with the aim to highlight which are the most important features. In particular, for all of the three classification models the U_o is the most impacting variable followed by U_w or R. The efficiency of the heating system and the heat transfer surface are always ranked as the 4^{th} and 5^{th} most impacting variable respectively, while

the other extensive geometric variables (i.e., V, A and H) have the lowest importance. This is consistent with the fact that the dataset includes only EPCs referred to flats that can be considered quite similar for what concerns heated gross volume, floor area and average ceiling height.

7.2. Clustering Results

After the identification of the best classifiers to be used as energy benchmarking models for the considered set of flats, the clustering stage was performed. In particular, among each pairs of contiguous energy performance classes, the K-means algorithm was used to identify the main groups of misclassified or correctly classified flats according to the eight input variables considered in the classification stage. As previously explained, the clustering algorithm was initialised setting the number of desired clusters K = 10 for each type of instances, then only the most populated clusters, that cover at least the 90% of total, were considered.

Instance type	**Cluster ID**	**n. of instances**	**% of total**
$A \rightarrow B$	1	607	53.20%
	2	414	36.28%
	3	89	7.80%
$B \rightarrow A$	1	422	43.59%
	2	296	30.60%
	3	154	15.86%
$A \rightarrow A$	1	2494	73.85%
	2	445	13.18%
	3	431	12.76%
$B \rightarrow B$	1	2544	74.13%
	2	539	15.71%
	3	232	6.76%

Table 4: Cardinality of clusters identified among correct and wrong classified instances pertaining to the contiguous energy performance classes A and B

Tab. 4 reports the obtained results related to the contiguous energy performance classes A and B. In particular, the tags $A \rightarrow B$, $B \rightarrow A$ refer to the flats misclassified in the class B and A respectively, while tags $A \rightarrow A$ and $B \rightarrow B$ were assigned to the flats correctly classified in A and B. From the table it can be seen that for each type of instances considered, the biggest three clusters were able to include more than the 90% of the instances. It

means that, according to the concept of similarity behind cluster analysis, by explaining the predictions performed for the centroids of the 12 clusters considered, the analyst can have a look at local behaviours of the classifier that can be used for extracting significant inference mechanisms learnt by the energy benchmarking model. In order to better characterise each cluster identified for the classes A and B, the components of each centroid were reported in Tab.5 with the evidence of the relative calculated average PED_h value.

Instance type	**Cluster ID**	A	V	H	S	R	U_o	U_w	η_h	PED_h
	1	80	305	3.8	213	0.7	0.5	2.5	0.80	79
$A \rightarrow B$	2	71	277	4.0	104	0.4	0.9	3.1	0.70	78
	3	185	688	3.8	460	0.7	0.5	2.3	0.80	73
	1	191	759	4.0	541	0.75	0.3	1.9	0.80	104
$B \rightarrow A$	2	68	259	3.8	93	0.4	0.7	2.6	0.80	108
	3	70	264	3.9	189	0.7	0.4	2.0	0.80	106
	1	72	277	3.9	156	0.6	0.4	1.9	0.80	59
$A \rightarrow A$	2	194	777	4.0	543	0.7	0.3	1.8	0.90	65
	3	67.5	265	3.9	79.4	0.3	0.8	2.8	0.80	62
	1	73	284	3.8	169	0.7	0.8	2.9	0.70	121
$B \rightarrow B$	2	62	232	4.0	164	0.3	0.85	4.9	0.70	116
	3	189	711	3.8	478	0.7	0.6	2.6	0.80	119

Table 5: Components of cluster centroids related to energy performance classes A and B

In addition Fig. 7 shows a graphical representation of the normalised centroid components referred to the most populated cluster of each instance type. The figure is particularly useful to infer some main differences among each group of instances. For example the centroids of the cluster 1 of $A \rightarrow A$ and $B \rightarrow B$ describe similarities in terms of geometric features (A, V, H, R) and dissimilarity for thermophysical properties, such as U_o and U_w. For what concerns the centroids of the cluster 1 of $A \rightarrow B$ and $B \rightarrow A$ the combination of the input variables leads to PED_h values that are close to the border (i.e., 91 kWh/m^2y) between the two contiguous energy performance classes A and B. As explained in the section 6 the same clustering process was also performed on the correct and wrong predictions of the classes $B-C$ and $C-D$. The main results were included in Appendix A and Appendix B.

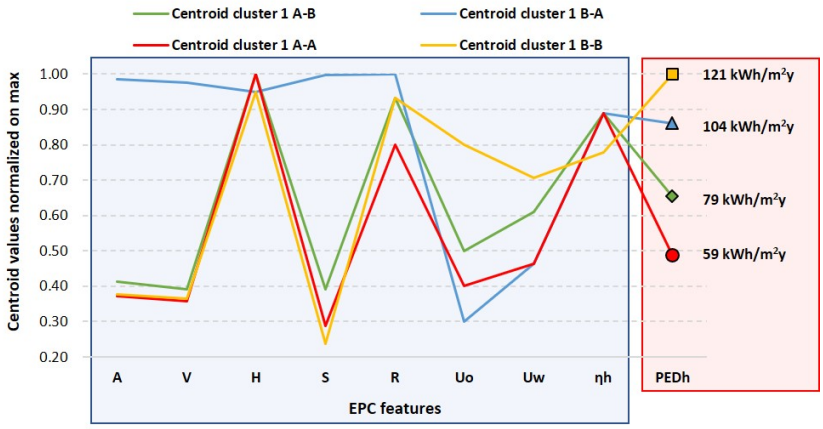
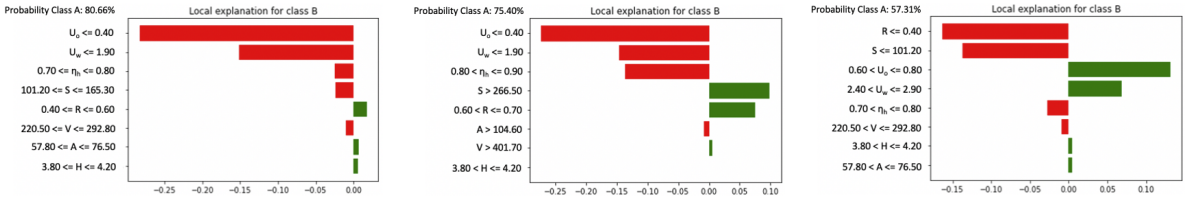


Figure 7: Graphical representation of the normalized centroid components related to the biggest cluster evaluated for each instance type.

7.3. Explanation Results

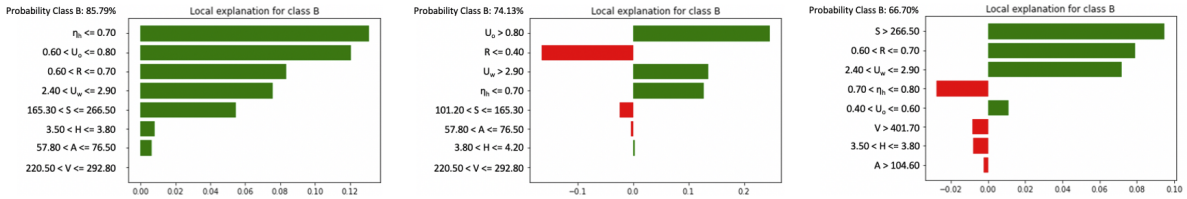
The results of the clustering analysis allowed to identify centroids that can be considered as archetypes among the analysed flats. The components of each centroid were used as input values of the classification model in order to predict for those objects the membership to a specific energy perfomance class. The next step employed the LIME tool for explaining why the model produced a specific prediction considering both correctly and wrongly classified flats.

In this section, for the sake of brevity, we only reported the explanation outputs related to the clusters of the $A-B$ classes. The results obtained for classes $B-C$ and $C-D$ were included in Appendix A and B respectively. In particular, as shown in Fig. 8, 9, 10 and 11, the bars of colour red indicate the variables (and their specific numeric ranges), that support the model in predicting the class A while the green bars indicate the features that had the opposite effect dragging the prediction toward class B. The combined effect of all the input variables determined the final probability class value that was reported in the top left corner of each figure. The first three explanations, pertaining the centroids representative of the flats labelled as A and correctly classified, are shown in Fig. 8. The explanations of the centroid 1 and 2 were characterized by a very high class probability (over 75%). It means that when the energy benchmaking model classified the archetype flats corresponding to the two centroids, the model expressed a very high confidence in estimating their membership to the energy performance class A.



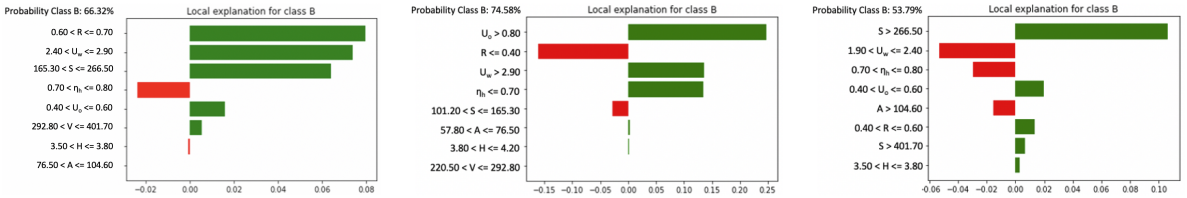
((a)) Centroid of cluster 1 ((b)) Centroid of cluster 2 ((c)) Centroid of cluster 3

Figure 8: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, evaluated among the instances $A \rightarrow A$



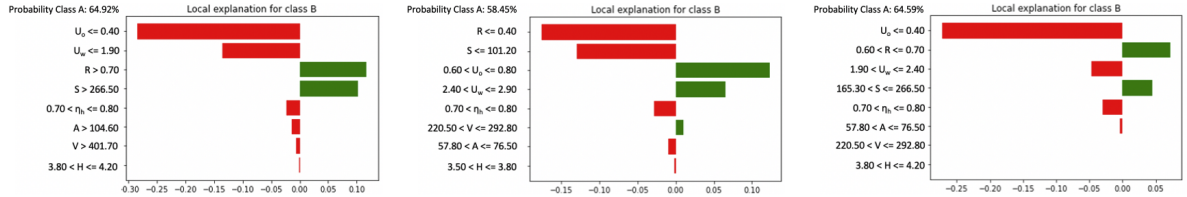
((a)) Centroid of cluster 1 ((b)) Centroid of cluster 2 ((c)) Centroid of cluster 3

Figure 9: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, evaluated among the instances $B \rightarrow B$



((a)) Centroid of cluster 1 ((b)) Centroid of cluster 2 ((c)) Centroid of cluster 3

Figure 10: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, evaluated among the instances $A \rightarrow B$



((a)) Centroid of cluster 1 ((b)) Centroid of cluster 2 ((c)) Centroid of cluster 3

Figure 11: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, evaluated among the instances $B \rightarrow A$

In particular the greatest impact toward class A is associated to the variables U_o and U_w that are lower than 0.4 and 1.9 W/m^2K respectively suggesting the presence of well insulated envelope. Despite this, the 3^{rd} centroid was correctly classified but with a low confidence (i.e., 57%) due to higher values of U_o and U_w (that suggested a classification in class B) but lower heat transfer surface S and aspect ratio R respect to the other two centroids. It means that flats with these characteristics, can lead to potential weak predictions of the energy benchmarking model even though correctly classified. Fig. 9 shows the explanation results pertaining the predictions of the three centroids evaluated for the instances correctly predicted as B. The 1^{st} centroid was classified with a very high confidence of about 85%. For this case, all the variables that had an impact on the explanation, supported the classification in the right class. In particular, the most important features ($\eta_h \leq 0.70$ and $0.60 < U_o \leq 0.80$) describe a flat with both envelope and heating system less efficient than flats in the energy performance class A. The 2^{nd} centroid was classifed with a confidence of about 75%. In this case, the two most important variables disagreed in the prediction explanation. In fact, a value of U_o higher than 0.80 W/m^2K supports the class B while the aspect ratio $R \leq 0.40$ is typical of flats in class A. A low value of R often corresponds to small areas of heat transfer surfaces S then giving less importance to envelope performance variables. The last centroids of the instances $B \rightarrow B$, was explained with a probability of 67% with the most three significant variables that support the class B.

Fig. 10 shows the explanations referred to the misclassified instances of the type $A \rightarrow B$. In particular, the 1^{st} centroid is wrongly classified as B with a strong confidence of about 67%. For this case the geometrical and thermophisical variables have values closer to the ones of class B rather then class A. Only the η_h is consistent with the global efficiency values of buildings labelled as A. The buildings with this specific configuration of variables are particularly difficult to be classified also considering that they are characterized by an annual PED_h (of 79 kWh/m^2y) value very close to the left border between the classes A and B (91 kWh/m^2y). Similarly, the 2^{nd} centroid was also misclassified with high probability (i.e., 74%) due to a particular combination of the building features. For this case, the high values of U_o and U_w were mitigated by a small heat transfer surface and low aspect ratio. It means that flats with these characteristics belong to class A, mostly due to geometrical aspects that positively impact on the heating energy need. Conversely, the 3^{rd} centroid was classified as B despite

the values of U_w and η_h were consistent with the ones of flats in class A. However the geometry of the flat (Heat transfer surface and aspect ratio) and an high value U_o have a negative impact on the transmission heat losses. The resuls is a misclassification with a very low probability for class B. It means that the classification of such instance among class A and B is almost random for the developed classifier.

Eventually, Fig. 11 shows the explanations pertaining to the misclassified instances of the type $B \rightarrow A$. Also in these cases the combination of high values of aspect ratios combined with low values of thermal transmittances and vice versa, represent the main source of misclassification.

As a final remark, all the performed explanations demonstrated to be strongly consistent with the results obtained through the global SHAP analysis (Fig. 6) in terms of feature importance.

8. Discussion

The present paper focused on the analysis of EPCs evaluated for about 100,000 flats located in Piedmont (North-western region of Italy). The proposed methodology was based on the analysis of open data of EPCs and provides a robust approach for the automatic asset rating of flat energy performance. The methodology proposes a classification approach to benchmark 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, the interpretation and the possible exploitation of the results obtained are discussed.

The classification process was based on the transformation of the numerical variable PED_h in four categorical contiguous classes identified according to the principle of equal frequency. This choice was driven by expert knowledge, in order to avoid class imbalance problems. However the ranges of PED_h that characterise each class resulted to be different between each other especially for what concern energy performance class labelled as D. The classification layer was designed to be flexible and generalizable as much as possible. In particular the classification was addressed as a three-step process. Firstly the EPC dataset was segmented in contiguous classes generating three binary datasets $A - B$, $B - C$ and $C - D$ respectively. Secondly, for each pair of classes five different classification models were trained and validated. Eventually, a model selection technique technique was implemented in order to combine the best models, which aimed to identify the right model

to use with unseen data of a real-world scenario. The proposed classification process outperformed the traditional multiclass approach by 5% in terms of accuracy. From a methodological perspective, the experimental evaluation demonstrated that the approach allows to produce differentiated models, able to fit better the specific features of the related EPC segments by exploiting a limited number of input variables. In fact, thanks to the adoption of the model selection technique it was possible to use, for the pairs of contiguous energy performance classes, different algorithms to address the classification task. In addition, it is worth to note that the proposed classification approach still remains valid also considering a different discretization of the target variable that can be then set by the user (e.g., public authority) according to its specific needs. Another innovative aspect of the present work, pertains to the introduction of a generalizable approach for explaining the estimation capabilities learnt by the classification models. The explanation layer makes the results obtained from the developed energy performance benchmarking model, understandable and exploitable also for non-domain experts. Useful information can be obtained from this benchmarking tool as it helps to discover in a straightforward way energy patterns among large dataset and at the same time understand the strengths and limitations of the estimation tool developed. To this purpose the agnostic XAI tool LIME was employed. LIME is a local explainer extremely effective in providing easy interpretations of classification results for binary problems. However, the local nature of the explanation provided should be taken into account for ensuring the feasibility of the prediction explanations, especially if large datasets are analysed. To overcome such barrier, a two-step analysis was proposed, combining LIME with an unsupervised clustering technique (i.e., K-means). The main reason behind this analysis, lied on the opportunity of extracting prototype instances in the dataset (i.e., cluster centroids) from specific groups of flats characterised by similar features (geometry, opaque and transparent envelope transmittances, system efficiencies etc.). In this way, leveraging the concept of similarity, exploited by clustering analysis, it was possible to provide local explanations of a representative combination of building features. This approach made it possible to better investigate rightly and wrongly classified instances understanding which are the main combinations of input variables that led to a specific classification results. In this perspective the explanation layer offers different opportunities from both analyst and end-user side (e.g., public authority, energy regulator, building portfolio manager). The advantages for the analyst can be summarized as

follows:

- break the trade-off between model complexity and model interpretability that often constrains energy benchmarking analysis;
- refine the model and improving the feature selection. The analyst can exploit the results of a XAI process for detecting the presence of input variables with low importance or which contribution to the learnt mechanism is meaningless and removing them from the input set;
- easily understand the strengths and weakness of the developed model. During the testing and validation of the model, the analyst can use the output of the XAI layer to infer specific patterns that can be associated with high/low performances of the model;
- infer the rational behind each prediction of the model.

On the other hand the advantages for the end-user can be summarized as follows:

- understand why a certain prediction is provided and what are the supporting and conflicting model features towards it;
- according to the explanations provided for each cluster centroid, identify which are the specific combinations of building features that can compromise the trustworthiness of the model;
- easily understand which are the most important features that strongly influence the energy performance of a flat/building.

As a reference, the end-user can be aware about specific feature combinations of flats that led the model, during the validation phase, to a misclassification. There is a great added value in this kind of information given that the end-user can associate to the prediction of an unknown instance (i.e., a new flat) an explanation and a misclassification risk according to which is the closest centroid considering the features of the flat. In general, designers and authority planners can exploit such tool to understand where put their effort, among large stocks of buildings, and which could be the most convenient retrofitting strategies to be promoted considering the feature combination of the buildings of interest. In this way it is possible to support the definition of robust financial investment policies that leverage such knowledge and help

to devise more targeted actions to improve energy efficiency in diversified stocks of buildings. Currently, XAI is then established as an essential requirement towards more effective and powerful AI-based systems in many domains, and energy is included as well. From a regularity perspective, XAI is also fundamental in order to verify machine learning models and to preserve characteristics such as fairness and transparency that are more and more required to modern AI-based decision support systems.

9. Conclusions

This paper proposed a multi-step methodology to automatically benchmark energy performance of flats by means of classification algorithms. The analysis was complemented including in the analytical process an explanation layer based on the LIME tool in order to make the results of the analysis interpretable as much as possible. The methodology was tested on a large collection of EPCs pertaining about 100,000 flats located in Piedmont (Italy).

Thanks to the performed analysis, also some limitations of this work have emerged, especially for what concerns the deployment phase of the energy performance benchmarking tool. In particular the main limitations of our work are related to the quality of EPCs data. In fact, EPCs can be subjected to different kinds of errors (i.e., clerical and calculation errors made by technicians while issuing the EPC) with potential negative impact on the accuracy of the developed models. In the perspective of increasing the size of the EPC source dataset, a robust pre-processing layer need to be deployed in order to avoid a decrease of model performance over time. In addition, still considering the deployment phase, the re-training of three binary classification models instead of one multiclass classifier, represents a task with higher computational cost.

Respect to the future work, further research will be focused on the testing of alternative configurations of algorithms (i.e., classifiers, clustering algorithms, XAI tools) with respect to the one considered in this study by performing a benchmark analysis between them especially from the explainability point of view. Eventually, particularly promising is the integration of a XAI layer also in other building Energy Management and Information System (EMIS) tools that typically leverage AI techniques such as advanced controllers and automated anomaly detection and diagnosis systems.

Acknowledgments

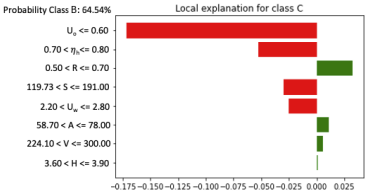
The authors express their gratitude to Giovanni Nuvoli (Settore Sviluppo Energetico Sostenibile Regione Piemonte) and to CSI Piemonte.

Appendix A

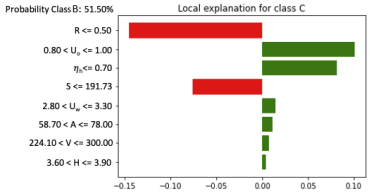
Appendix A reports the results related to the contiguous energy performance classes B and C. In particular, in Tab. A.1 the tags $B \rightarrow C$, $C \rightarrow B$ refer to the flats misclassified in the class B and C respectively, while tags $B \rightarrow B$ and $C \rightarrow C$ were assigned to the flats correctly classified in B and C. Successively, are reported the explanation outputs related to the clusters of the $B - C$ classes. In particular, as shown in Fig. A.1, A.2, A.3 and A.4, the bars of colour red indicate the variables (and their specific numeric ranges), that support the model in predicting the class B while the green bars indicate the features that had the opposite effect dragging the prediction toward class C. The combined effect of all the input variables determined the final probability class value that was reported in the top left corner of each figure.

Instance type	**Cluster ID**	**n. of instances**	**% of total**
$B \rightarrow C$	1	958	66.34%
	2	302	20.90%
	3	88	6.10%
$C \rightarrow B$	1	612	48.22%
	2	525	41.37%
	3	81	6.40%
$B \rightarrow B$	1	1606	52.63%
	2	996	32.64%
	3	392	12.84%
$C \rightarrow C$	1	2043	64.59%
	2	799	25.26%
	3	294	9.30%

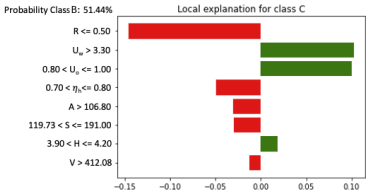
Table A.1: Cardinality of clusters identified among correct and wrong classifications of the contiguous energy performance classes B and C



((a)) Centroid of cluster 1

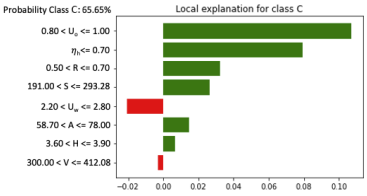


((b)) Centroid of cluster 2

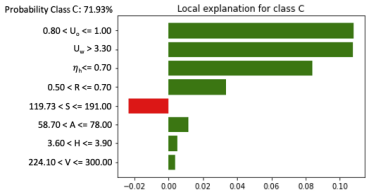


((c)) Centroid of cluster 3

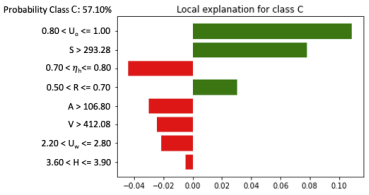
Figure A.1: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, evaluated among the instances $B \rightarrow B$



((a)) Centroid of cluster 1

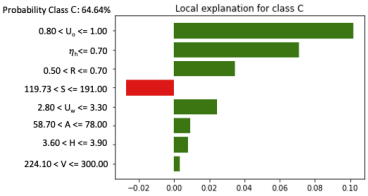


((b)) Centroid of cluster 2

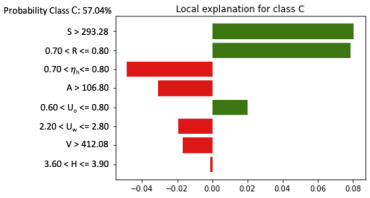


((c)) Centroid of cluster 3

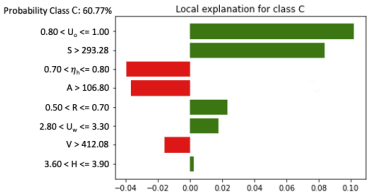
Figure A.2: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, evaluated among the instances $C \rightarrow C$



((a)) Centroid of cluster 1

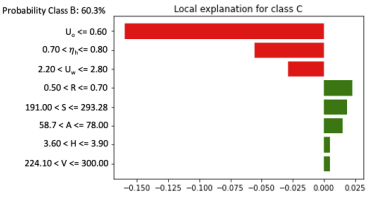


((b)) Centroid of cluster 2

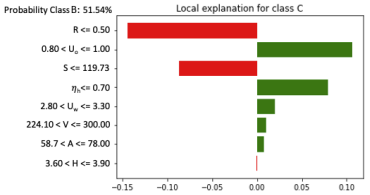


((c)) Centroid of cluster 3

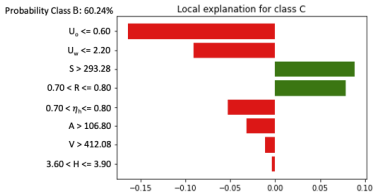
Figure A.3: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, evaluated among the instances $B \rightarrow C$



((a)) Centroid of cluster 1



((b)) Centroid of cluster 2



((c)) Centroid of cluster 3

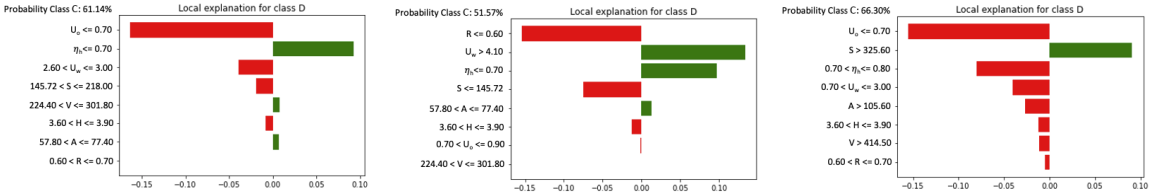
Figure A.4: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, evaluated among the instances $C \rightarrow B$

Appendix B

Appendix B reports the results related to the contiguous energy performance classes C and D. In particular, in Tab. B.1 the tags $C \rightarrow D$, $D \rightarrow C$ refer to the flats misclassified in the class C and D respectively, while tags $C \rightarrow C$and $D \rightarrow D$ were assigned to the flats correctly classified in C and D. Successively, are reported the explanation outputs related to the clusters of the $C - D$ classes. In particular, as shown in Fig. B.1, B.2, B.3 and B.4, the bars of colour red indicate the variables (and their specific numeric ranges), that support the model in predicting the class C while the green bars indicate the features that had the opposite effect dragging the prediction toward class D. The combined effect of all the input variables determined the final probability class value that was reported in the top left corner of each figure.

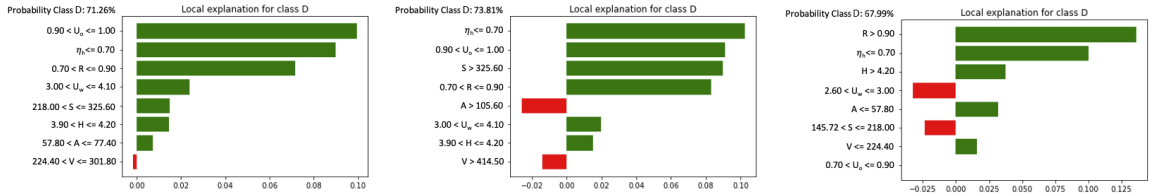
Instance type	**Cluster ID**	**n. of instances**	**% of total**
$C \rightarrow D$	1	996	81.77%
	2	192	15.76%
	3	8	0.70%
$D \rightarrow C$	1	810	58.95%
	2	327	23.80%
	3	112	8.15%
$C \rightarrow C$	1	2186	65.88%
	2	636	19.16%
	3	469	14.13%
$D \rightarrow D$	1	2676	88.25%
	2	245	8.10%
	3	57	1.90%

Table B.1: Cardinality of clusters identified among correct and wrong classifications of the contiguous energy performance classes C and D



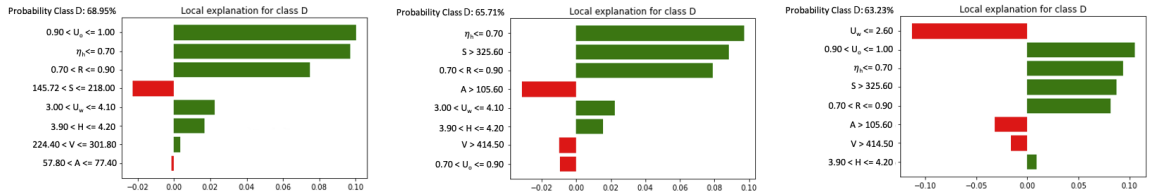
((a)) Centroid of cluster 1 ((b)) Centroid of cluster 2 ((c)) Centroid of cluster 3

Figure B.1: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, evaluated among the instances $C \rightarrow C$



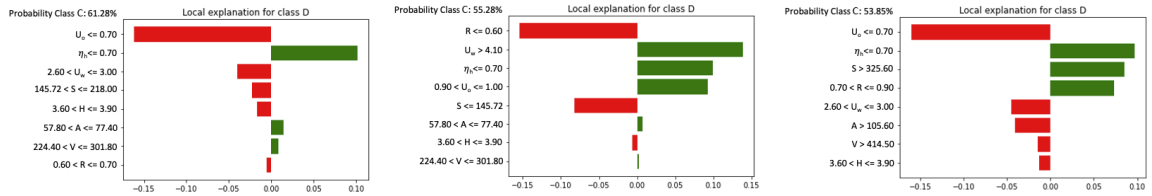
((a)) Centroid of cluster 1 ((b)) Centroid of cluster 2 ((c)) Centroid of cluster 3

Figure B.2: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, evaluated among the instances $D \rightarrow D$



((a)) Centroid of cluster 1 ((b)) Centroid of cluster 2 ((c)) Centroid of cluster 3

Figure B.3: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, among the instances $C \rightarrow D$



((a)) Centroid of cluster 1 ((b)) Centroid of cluster 2 ((c)) Centroid of cluster 3

Figure B.4: LIME outputs referred to the predictions of (a) centroid of Cluster 1, (b) centroid of cluster 2 and (c) centroid of cluster 3, evaluated among the instances $D \rightarrow C$

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138–52160. doi:10.1109/ACCESS.2018.2870052.

Akhlaghi, Y. G., Aslansefat, K., Zhao, X., Sadati, S., Badiei, A., Xiao, X., Shittu, S., Fan, Y., & Ma, X. (2020). Hourly performance forecast of a dew point cooler using explainable artificial intelligence and evolutionary optimisations by 2050. *Applied Energy*, *281*, 116062. doi:10.1016/j.apenergy.2020.116062.

Al-Homoud, M. S. (2001). Computer-aided building energy analysis techniques. *Building and Environment*, *36*, 421–433. doi:10.1016/S0360-1323(00)00026-3.

Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, *81*, 1192–1205. doi:10.1016/j.rser.2017.04.095.

Arjunan, P., Poolla, K., & Miller, C. (2020). Energystar++: Towards more accurate and explanatory building energy benchmarking. *Applied Energy*, *276*, 115413. doi:10.1016/j.apenergy.2020.115413.

Arjunan, P., Poolla, K., & Miller, C. (2022). Beem: Data-driven building energy benchmarking for singapore. *Energy and Buildings*, *260*, 111869. doi:10.1016/j.enbuild.2022.111869.

Attanasio, A., Piscitelli, M. S., Chiusano, S., Capozzoli, A., & Cerquitelli, T. (2019). Towards an automated, fast and interpretable estimation model of heating energy demand: A data-driven approach exploiting building energy certificates. *Energies*, *12*, 1273. doi:10.3390/en12071273.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & MÃžller, K.-R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, *11*, 1803–1831.

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)* (pp. 8–13). volume 8.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123–140. doi:10.1007/BF00058655.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32. doi:10.1023/A:1010933404324.

Capozzoli, A., Grassi, D., & Causone, F. (2015). Estimation models of heating energy consumption in schools for local authorities planning. *Energy and Buildings*, *105*, 302–313. doi:10.1016/j.enbuild.2015.07.024.

Capozzoli, A., Piscitelli, M. S., Neri, F., Grassi, D., & Serale, G. (2016a). A novel methodology for energy performance benchmarking of buildings by means of linear mixed effect model: The case of space and dhw heating of out-patient healthcare centres. *Applied Energy*, *171*, 592–607. doi:10.1016/j.apenergy.2016.03.083.

Capozzoli, A., Serale, G., Piscitelli, M. S., & Grassi, D. (2016b). Data mining for energy analysis of a large data set of flats. In *Proceedings of the Institution of Civil Engineers-Engineering Sustainability* (pp. 3–18). Thomas Telford Ltd volume 170. doi:10.1680/jensu.15.00051.

Chakraborty, D., Başağaoğlu, H., & Winterle, J. (2021). Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Systems with Applications*, *170*, 114498. doi:10.1016/j.eswa.2020.114498.

Chen, H.-Y., & Lee, C.-H. (2020). Vibration signals analysis by explainable artificial intelligence (xai) approach: Application on bearing faults diagnosis. *IEEE Access*, *8*, 134246–134256. doi:10.1109/ACCESS.2020.3006491.

Chung, W. (2011). Review of building energy-use performance benchmarking methodologies. *Applied Energy*, *88*, 1470–1479. doi:10.1016/j.apenergy.2010.11.022.

Di Corso, E., Cerquitelli, T., Piscitelli, M. S., & Capozzoli, A. (2017). Exploring energy certificates of buildings through unsupervised data mining techniques. In *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)* (pp. 991–998). IEEE. doi:10.1109/iThings-GreenCom-CPSCom-SmartData.2017.15.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210–0215). IEEE.

Eskandarnia, E., Al-Ammal, H. M., & Ksantini, R. (2022). An embedded deep-clustering-based load profiling framework. *Sustainable Cities and Society*, *78*, 103618. doi:10.1016/j.scs.2021.103618.

Fabrizio, E., Corrado, V., & Filippi, M. (2010). A model to design and optimize multi-energy systems in buildings at the design concept stage. *Renewable Energy*, *35*, 644–655. doi:10.1016/j.renene.2009.08.012.

Fan, C., Sun, Y., Zhao, Y., Song, M., & Wang, J. (2019a). Deep learning-based feature engineering methods for improved building energy prediction. *Applied energy*, *240*, 35–45. doi:10.1016/j.apenergy.2019.02.052.

Fan, C., Xiao, F., Yan, C., Liu, C., Li, Z., & Wang, J. (2019b). A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. *Applied Energy*, *235*, 1551–1560. doi:10.1016/j.apenergy.2018.11.081.

Filogamo, L., Peri, G., Rizzo, G., & Giaccone, A. (2014). On the classification of large residential buildings stocks by sample typologies for energy planning purposes. *Applied Energy*, *135*, 825–835. doi:10.1016/j.apenergy.2014.04.002.

Frick, N. M., Schiller, S., Stuart, E., Schwartz, L., Kramer, C., & Faesy, R. (2017). *Evaluation of U.S. Building Energy Benchmarking and Transparency Programs: Attributes, Impacts, and Best Practices*. Technical Report.

Galli, A., Moscato, V., Sperlí, G., & De Santo, A. (2020). An explainable artificial intelligence methodology for hard disk fault prediction. In *International Conference on Database and Expert Systems Applications* (pp. 403–413). Springer. doi:10.1007/978-3-030-59003-1_26.

Gao, X., & Malkawi, A. (2014). A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm. *Energy and Buildings*, *84*, 607–616. doi:10.1016/j.enbuild.2014.08.030.

Geraldi, M. S., & Ghisi, E. (2022). Data-driven framework towards realistic bottom-up energy benchmarking using an artificial neural network. *Applied Energy*, *306*, 117960. doi:10.1016/j.apenergy.2021.117960.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*, 3–42. doi:10.1007/s10994-006-6226-1.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*, 1–42. doi:10.1145/3236009.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, *2*.

Himeur, Y., Ghanem, K., Alsalemi, A., Bensaali, F., & Amira, A. (2021). Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Applied Energy*, *287*, 116601. doi:10.1016/j.apenergy.2021.116601.

Hong, T., Piette, M. A., Chen, Y., Lee, S. H., Taylor-Lange, S. C., Zhang, R., Sun, K., & Price, P. (2015). Commercial building energy saver: an energy retrofit analysis toolkit. *Applied Energy*, *159*, 298–309. doi:10.1016/j.apenergy.2015.09.002.

ISO-15927 (2003). Iso 15927-1, hygrothermal performance of buildings-calculation and presentation of climatic data - part 1: Monthly means of single meteorological elements. *International Organization for Standardization*, . URL: https://www.iso.org/standard/29559.html.

Kuzlu, M., Cali, U., Sharma, V., & Güler, Ö. (2020). Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. *IEEE Access*, *8*, 187814–187823. doi:10.1109/ACCESS.2020.3031477.

Lee, S. H., Hong, T., Piette, M. A., & Taylor-Lange, S. C. (2015). Energy retrofit analysis toolkits for commercial buildings: A review. *Energy*, *89*, 1087–1100. doi:10.1016/j.energy.2015.06.112.

Lee, W., Yik, F., & Burnett, J. (2003). A method to assess the energy performance of existing commercial complexes. *Indoor and Built Environment*, *12*, 311–327. doi:10.1177/142032603035424.

Lee, W.-S., & Lee, K.-P. (2009). Benchmarking the performance of building energy management using data envelopment analysis. *Applied Thermal Engineering*, *29*, 3269–3273. doi:10.1016/j.applthermaleng.2008.02.034.

Li, K., Sun, Y., Robinson, D., Ma, J., & Ma, Z. (2020). A new strategy to benchmark and evaluate building electricity usage using multiple data mining technologies. *Sustainable Energy Technologies and Assessments*, *40*, 100770. doi:10.1016/j.seta.2020.100770.

Li, Z., Han, Y., & Xu, P. (2014). Methods for benchmarking building energy consumption against its past or intended performance: An overview. *Applied Energy*, *124*, 325–334. doi:10.1016/j.apenergy.2014.03.020.

Liu, X., Ding, Y., Tang, H., & Xiao, F. (2021). A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. *Energy and Buildings*, *231*, 110601. doi:10.1016/j.enbuild.2020.110601.

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint*, (pp. 1–223). doi:10.48550/ARXIV.1407.7502.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777). doi:10.48550/arXiv.1705.07874.

Madhikermi, M., Malhi, A. K., & Främling, K. (2019). Explainable artificial intelligence based heat recycler fault detection in air handling unit. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (pp. 110–125). Springer. doi:10.1007/978-3-030-30391-4_7.

Mauro, G. M., Hamdy, M., Vanoli, G. P., Bianco, N., & Hensen, J. L. (2015). A new methodology for investigating the cost-optimality of energy retrofitting a building category. *Energy and Buildings*, *107*, 456–478. doi:10.1016/j.enbuild.2015.08.044.

Millar, D., Tonolo, G., & Ziebinska, U. (2016). Energy efficiency indicators: Highlights. *International Energy Agency*

(IEA), Paris, France, . URL: https://www.iea.org/reports/energy-efficiency-indicators-overview.

Miller, C. (2019a). More buildings make more generalizable models—benchmarking prediction methods on open electrical meter data. *Machine Learning and Knowledge Extraction*, *1*, 974–993. doi:10.3390/make1030056.

Miller, C. (2019b). What's in the box?! towards explainable machine learning applied to non-residential building smart meter classification. *Energy and Buildings*, *199*, 523–536. doi:10.1016/J.ENBUILD.2019.07.019.

Papadopoulos, S., & Kontokosta, C. E. (2019). Grading buildings on energy performance using city benchmarking data. *Applied Energy*, *233*, 244–253. doi:10.1016/j.apenergy.2018.10.053.

Pasichnyi, O., Wallin, J., Levihn, F., Shahrokni, H., & Kordas, O. (2019). Energy performance certificates—new opportunities for data-enabled urban energy policy instruments? *Energy Policy*, *127*, 486–499. doi:10.1016/j.enpol.2018.11.051.

Pereira, S., Meier, R., Alves, V., Reyes, M., & Silva, C. A. (2018). Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment. In *Understanding and interpreting machine learning in medical image computing applications* (pp. 106–114). Springer. doi:10.1007/978-3-030-02628-8_12.

Petcharat, S., Chungpaibulpatana, S., & Rakkwamsuk, P. (2012). Assessment of potential energy saving using cluster analysis: A case study of lighting systems in buildings. *Energy and Buildings*, *52*, 145–152. doi:10.1016/j.enbuild.2012.06.006.

Piscitelli, M. S., Mazzarelli, D. M., & Capozzoli, A. (2020). Enhancing operational performance of ahus through an advanced fault detection and diagnosis process based on temporal association and decision rules. *Energy and Buildings*, *226*, 110369. doi:10.1016/j.enbuild.2020.110369.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81–106. doi:10.1007/BF00116251.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). doi:10.48550/arXiv.1602.04938.

Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, *53*, 23–69. doi:10.1023/A:1025667309714.

Roth, J., Lim, B., Jain, R. K., & Grueneich, D. (2020). Examining the feasibility of using open data to benchmark building energy usage in cities: A data science and policy perspective. *Energy Policy*, *139*, 111327. doi:10.1016/j.enpol.2020.111327.

Runge, J., & Zmeureanu, R. (2019). Forecasting energy use in buildings using artificial neural networks: A review. *Energies*, *12*, 3254. doi:10.3390/en12173254.

Sardianos, C., Varlamis, I., Chronis, C., Dimitrakopoulos, G., Alsalemi, A., Himeur, Y., Bensaali, F., & Amira, A. (2021). The emergence of explainability of intelligent systems: Delivering explainable and personalized recommendations for energy efficiency. *International Journal of Intelligent Systems*, *36*, 656–680. doi:10.1002/INT.22314.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626). doi:10.1109/ICCV.2017.74.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint*, (pp. 1–14). doi:10.48550/arXiv.1412.6806.

Sun, Y., Haghighat, F., & Fung, B. C. (2020). A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy and Buildings*, *221*, 110022. doi:10.1016/j.enbuild.2020.110022.

Tahsildoost, M., & Zomorodian, Z. S. (2015). Energy retrofit techniques: An experimental study of two typical school buildings in tehran. *Energy and Buildings*, *104*, 65–72. doi:10.1016/j.enbuild.2015.06.079.

Tan, P.-N., Steinbach, M., & Kumar, V. (2013). Data mining cluster analysis: basic concepts and algorithms. *Introduction to data mining*, (pp. 487–533).

UNI-10349 (2016). Uni 10349-1:2016, heating and cooling of buildings - climatic data - part 1: Monthly means for evaluation of energy need for space heating and cooling and methods for splitting global solar irradiance into the direct and diffuse parts and for calculate the solar irradiance on tilted planes. *Italian Organization for Standardization*, . URL: `http://store.uni.com/`.

UNI/TS-11300 (2008). Uni/ts 11300-2, energy performance of buildings-part 2. *Italian Organization for Standardization*, . URL: `https://www.gazzettaufficiale.it/eli/id/2009/07/10/09A07900/sg`.

Wang, S., Yan, C., & Xiao, F. (2012). Quantitative energy performance assessment methods for existing buildings. *Energy and buildings*, *55*, 873–888. doi:`10.1016/j.enbuild.2012.08.037`.

Yang, Z., Roth, J., & Jain, R. K. (2018). Due-b: Data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis. *Energy and Buildings*, *163*, 58–69. doi:`10.1016/j.enbuild.2017.12.040`.

Zhang, Y., O'Neill, Z., Dong, B., & Augenbroe, G. (2015). Comparisons of inverse modeling approaches for predicting building energy performance. *Building and Environment*, *86*, 177–190. doi:`10.1016/j.buildenv.2014.12.023`.

Zhao, Y., Zhang, C., Zhang, Y., Wang, Z., & Li, J. (2020). A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. *Energy and Built Environment*, *1*, 149–164. doi:`10.1016/j.enbenv.2019.11.003`.