Optimizing building retrofit through data analytics: A study of multi-objective optimization and surrogate models derived from energy performance certificates

Article *in* Energy and Built Environment · July 2023

DOI: 10.1016/j.enbenv.2023.07.002

4 authors:

Gonçalo Roque Araújo
Technical University of Lisbon

Ricardo A. Gomes
Technical University of Lisbon

Paulo Ferrão
Instituto Superior Técnico

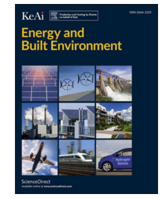
M. Glória Gomes
University of Lisbon

Energy and Built Environment xxx (xxxx) xxx

Contents lists available at ScienceDirect

Energy and Built Environment

journal homepage: http://www.keaipublishing.com/en/journals/energy-and-built-environment/



Optimizing building retrofit through data analytics: A study of multi-objective optimization and surrogate models derived from energy performance certificates

G.R. Araújo [a,b,*], Ricardo Gomes [a], Paulo Ferrão [a], M. Glória Gomes [b]

[a] IN+, DEM, Técnico Lisboa, Universidade de lisboa, Av Rovisco Pais, Lisboa, 1049-001, Portugal
[b] CERIS, DECivil, Técnico Lisboa, Universidade de Lisboa, Av Rovisco Pais, Lisboa, 1049-001, Portugal

ARTICLE INFO

Keywords:
Building energy performance
Building optimization
Multi-Objective surrogate models
Building retrofitting

ABSTRACT

The building stock is responsible for a large share of global energy consumption and greenhouse gas emissions, therefore, it is critical to promote building retrofit to achieve the proposed carbon and energy neutrality goals. One of the policies implemented in recent years was the Energy Performance Certificate (EPC) policy, which proposes building stock benchmarking to identify buildings that require rehabilitation. However, research shows that these mechanisms fail to engage stakeholders in the retrofit process because it is widely seen as a mandatory and complex bureaucracy. This study makes use of an EPC database to integrate machine learning techniques with multi-objective optimization and develop an interface capable of (1) predicting a building's, or household's, energy needs; and (2) providing the user with optimum retrofit solutions, costs, and return on investment. The goal is to provide an open-source, easy-to-use interface that guides the user in the building retrofit process. The energy and EPC prediction models show a coefficient of determination (R^2) of 0.84 and 0.79, and the optimization results for one case study EPC with a 2000€ budget limit in Évora, Portugal, show decreases of up to 60% in energy needs and return on investments of up to 7 in 3 years.

1. Introduction

One of the largest shares of energy consumption is associated with the building stock (in Europe, it is responsible for 40% of the total energy consumption), and since a building's lifetime can exceed 100 years it is important to improve energy regulation and develop instruments that promote the reduction of greenhouse gas emissions without neglecting thermal comfort and quality of life of its occupants [1]. Pasichnyi et al. [2] and Pérez-Lombard et al. [3] point to regulation, audits, and certification as three basic policy instruments for enhancing energy efficiency in buildings. In this context, building Energy Performance Certificates (EPCs) emerged to help achieve energy efficiency in buildings since the early 1990s [2], and their main goals are threefold: (1) to inform stakeholders of the building sector about building energy consumption and performance [4], (2) to decarbonize the building stock, and (3) to enhance investment in more efficient and sustainable solutions and systems, as documented on the Building Directive update from 2018 [5].

Building energy certification has been developed as a key policy instrument to improve buildings' energy efficiency, decrease energy consumption, and provide more transparency on energy use in buildings. An energy performance certificate (EPC) is a central element of the Energy Performance Building Directive (EPBD). In 2002, the EU Parliament defined EPC as " a document recognized by a Member State or by a legal person designated by it, which indicates the energy performance of a building or building unit" [5,6].

Regarding building energy certification, Pérez-Lombard et al. [3] distinguish three main advantages: (i) bench-marking, (ii) rating, and (iii) labeling. Nikolaou et al. [7] add (iv) building stock databases and methods for improving energy efficiency. Pasichnyi et al. [2] reviewed 79 papers focused on EPC applications from data analyses and stated that most EPC data have wider applications than initially intended by the EPC policy instrument. The detailed characterization of the building stock provided by the EPC scheme can be applied in the development of tools and applications enhancing energy efficiency measures.

Alongside, energy-efficient building retrofit is a key aspect to reduce carbon emissions [8], improving public health [9], and creating new jobs [10]. Despite their multiple benefits and efforts to promote building retrofits, the retrofit rates worldwide remain low, usually less than 1% per year [11]. One crucial barrier to this low number is the lack of knowledge of which combinations of retrofits are most cost-effective.

* Corresponding author.
E-mail address: goncalo.r.araujo@tecnico.ulisboa.pt (G.R. Araújo).

https://doi.org/10.1016/j.enbenv.2023.07.002
Received 19 April 2023; Received in revised form 3 July 2023; Accepted 8 July 2023
Available online xxx

Please cite this article as: G.R. Araújo, R. Gomes, P. Ferrão et al., Optimizing building retrofit through data analytics: A study of multi-objective optimization and surrogate models derived from energy performance certificates, Energy and Built Environment, https://doi.org/10.1016/j.enbenv.2023.07.002

Table 1
EPCs' energy indicators.

Indicators	Description
Nic	Annual Nominal needs of useful energy for heating (kWh/m^2)
Ni	Annual Reference needs of useful energy for heating (kWh/m^2)
Nvc	Annual Nominal needs of useful energy for cooling (kWh/m^2)
Nv	Annual Reference needs of useful energy for cooling (kWh/m^2)
Nac	Annual Nominal needs of useful energy for the production of domestic hot water (kWh/m^2)
Na	Annual Reference needs of useful energy for the production of domestic hot water (kWh/m^2)
Ntc	Annual Total nominal primary energy needs (kWh/m^2)
Nt	Annual Reference primary energy needs (kWh/m^2)

Table 2
EPCs' energy efficiency labels.

Energy label	$R = Ntc/Nt$
A+	$R \leq 0.25$
A	$0.25 < R \leq 0.5$
B	$0.5 < R \leq 0.75$
B-	$0.75 < R \leq 1$
C	$1 < R \leq 1.5$
D	$1.5 < R \leq 2$
E	$2 < R \leq 2.5$
F	$2.5 < R \leq 3$

1.1. Background

A household's EPC is issued by qualified experts (architects or engineers) who audit the households and collect information regarding building geometry, constructive solutions, heating, ventilation and air-conditioning (HVAC) systems, domestic water heating systems (DHW), and energy production systems, to determine the energy performance parameters and calculate energy performance indicators (including the energy label). Moreover, improvement measures and their impact on the energy label and energy consumption are suggested on the EPC by the experts [12].

The EPC considers numerical calculation methods based on the ISO standards (ISO 52000-1 [13], 52003-1 [14], 52010-1 [15], 52016-1 [16], and 52018-1 [17]) for heating, cooling, and domestic hot water needs and provide energy efficiency benchmark by considering reference values [18,19]. The energy needs indicators calculated for the energy certification process are summarized in Tables 1 and 2.

In most European countries, the energy efficiency labeling of a household assumes eight classes (from A+, the most energy-efficient to F, the least) [5] and is calculated by framing the parameter $R = Ntc/Nt$ in specific range values (Table 2).

The building EPC benchmarking system and its related renovation policies have significantly contributed to the improvement of buildings' energy use and to achieving the European renovation wave goals [20,21]. However, they still present constraints that are preventing citizens and stakeholders from implementing building retrofits such as a general lack of engagement and EPC data errors.

The lack of awareness and engagement with the EPC system is discussed by Watts et al. [22], which surveyed how people perceive EPCs, their importance, and if they implement the suggested retrofits. Results show that the EPC scheme is perceived more as mandatory bureaucracy and most people do not follow retrofit indications. In addition to this, Hardy & Glew [23] showed that 36% to 62% of EPCs in the United Kingdom contain errors and that these are mostly caused by EPC assessors. Consequently, the authors consider that new methodologies can be developed to prevent misleading results and reduce input complexity, and conclude that ML has great potential to help tackle these issues. Thus, the use of ML techniques to create models capable of predicting a building EPC or energy needs with easy-to-grasp inputs can help in: (1) identifying, controlling, and correcting issued EPCs with input errors [24]; and (2) engaging the community with the EPC scheme and its benefits [22].

Most studies regarding EPC surrogate models found in the literature focus on accurately predicting and validating existing EPCs. Buratti et al. [25] used an Artificial Neural Networks (ANN) model developed with a database of 6500 EPCs received by the Umbria Region (central Italy). The developed model allowed the authors to evaluate regional building energy consumption. Furthermore, the authors evaluated the accuracy of the model and identified EPCs requiring data correction. To engage the community with the EPC scheme, Khayatian et al. [26] simplify the EPCs inputs parameters and features, by developing a regression model using ANN, with acceptable accuracy and capable of predicting building energy consumption and EPC results with only 12 features.

Such models can also be useful to improve households' energy use and perform the best possible retrofits according to their specific needs and capabilities, by both reducing input complexity and allowing the application of optimization techniques (2). In this sense, Fan & Xia [27] developed two optimization models to find the best building retrofit that yielded maximum energy savings and minimum investment payback time. This is done considering the South African building EPC scheme and its tax incentive initiative program.

Multi-Objective Optimization (MOO) is often used when dealing with complex systems such as building retrofits since multiple conflicting objectives are typically entailed (e.g., better-performing buildings entail higher retrofit costs) [28]. When dealing with multiple objectives, optimum solutions are described as non-dominated solutions, which cannot improve in one objective, without harming the other [29]. Because most building retrofit optimization objectives are outputs either from time-consuming simulation or surrogate models and typically entail multiple objectives, they are typically solved with metaheuristics [30].

1.2. Research gap and goals

As stated in the Renovation Wave program [20], it is of critical importance to perform building retrofits as a way to reduce building energy consumption. In spite of multiple policies deployed to increase the renovation wave, there are still some barriers preventing its success such as insufficient information on the current energy profile of buildings and policies [22], lack of trust in the actual energy savings estimation [11], difficult decision-making processes [31], and financial obstacles [8]. In this sense, it is crucial to provide citizens and stakeholders with valuable, accurate, and comprehensible information regarding their house or building energy performance and the benefits of specific retrofit measures. The integration of MOO with EPC surrogate models enables the development of easy-to-use interfaces capable of optimizing any household or building retrofit solution. This gives users valuable insight into their EPC applicable renovation policies and government funding without extensive knowledge and bureaucracy.

The goal of this study is to strengthen the commitment to the renovation wave program [20] by developing an interface capable of quickly predicting and optimizing EPC indicators with easy-to-grasp inputs. Machine Learning techniques are applied to an EPC database of residential buildings in Portugal to develop regression models capable of predicting a building or household's annual energy needs, and EPC labels. Afterward, these models are integrated with a Multi-Objective Optimization (MOO) process to provide the user with a range of optimum retrofit solutions while estimating their impact on energy consumption and return on investment. This approach can be adapted to other countries' databases and yield information regarding energy performances and how to manage retrofit and rehabilitation investments at a larger scale. Moreover, researchers are provided with alternative approaches for building retrofit and optimization that require less computational power and have fewer inputs than building performance simulation and optimization using complex models. Finally, these techniques can be applied to the development of tools or applications to enhance the retrofit process by providing fast but accurate analysis of retrofit alternatives.

2. Methodology

This research proposes the use of ML techniques with Python programming language and the Sci-Kit Learn package [32] to develop an interface capable of engaging citizens and stakeholders in retrofitting their households and buildings by presenting the EPC prediction and optimization of the building retrofit process for a given budget. This includes the optimization of retrofit solutions for different objectives (energy and financial) by considering significantly fewer inputs than the energy certification process or building performance simulation. The methodology followed in this work can be segmented into three sub-sections: (1) Model development; (2) Multi-Objective Optimization (MOO); and (3) User interface - Case study (Fig. 1).

Sub-section (1) describes the development of regression models with supervised learning algorithms. These models predict the EPC energy indicators and the EPC label (Tables 1 and 2, Section 1.1). Initially, the first step is to clean the entire database by filtering noisy and inaccurate EPC data such as the date of issue, non-residential households, and unrealistic values of EPCs' input features. Secondly, a feature selection technique is applied to select the most representative features of the EPC label values. The final database is split into a training and test set that is used to train, test, and compare the performance of three regression algorithms by measuring their Root Mean Squared Error (RMSE) and coefficient of determination (R^2). A K-Fold cross-validation [33] is performed on the selected final model.

Sub-section (2) comprises the integration of the ML model in a Multi-Objective Optimization (MOO) process to (1) minimize an EPC's total annual nominal primary energy needs - *Ntc* $[kWh/m^2]$;(2) maximize the corresponding energy and tax savings (Return on Investment (ROI), i.e. the ratio between investment benefits and the cost of the investment); and (3) minimize the building retrofit cost.

Sub-section (3) focuses on interface development and its illustration in a case study. User inputs are defined, as well as analysis and visualization methods. The interface requires that users collect and introduce a different set of inputs representative of their household or building. Afterward, users obtain an estimation of their *Nic, Nvc, Ntc*, and EPC label (Tables 1 and 2, Section 1.1) while being able to visualize their results interactively. The integrated optimization process allows users to find the best retrofit solutions given a maximum retrofit budget and explore the results for different retrofit options.

2.1. Model development

For database preprocessing, the first step consists of filtering unrealistic values and splitting the resulting database into a training and test set for all variables and outputs. Then, feature selection techniques are applied to retrieve the most explainable features of the model [34]. For the models' training, the original training set is split into training and test sets in which an Extra Trees (ET) ensemble algorithm [35], a Multilayer Perceptron (MLP) ANN [36,37], and a Gradient Boosting (GB) for the regression problems [38] were used. These models have been extensively used in the literature for regression problems [39], particularly in building energy-related problems [40].

Table 3
EPCs' database feature list for k-best = 20.

General details	Construction elements	Equipment	Glazing
Property	Wall type	DHW source	Window area $[m^2]$
Year	Wall area $[m^2]$	Heating source	Window type
Area $[m^2]$	Roof type	DHW type	
Height [m]	Roof area $[m^2]$	Heating type	
Typology	Floor type	N° DHW equip	
N° floors		N° heating equip	
District			

Table 4
Test set target values distribution and indicators.

	Nic $[kWh/m^2]$	Nvc $[kWh/m^2]$	Ntc $[kWh/m^2]$	R [ratio]
count	5914	5914	5914	5914
mean	10.46	1.90	96.10	0.72
std	7.03	1.70	86.14	0.51
min	0.00	0.00	0.00	0.00
max	48.79	7.00	660.63	4.20

For the development of the ML models, the Portuguese EPC database is used, which contains over 800,000 certificates, and 88 features split between household details, opaque and glazed envelopes, and systems. From the whole database, all entries that were not issued under the 2006 Decree-Law [41] were removed, since the method of calculation of the EPC changed significantly due to the EPBD [6]. Additionally, all the EPCs without DHW equipment, all features with unrealistic values, and samples with over 70% of null values for all features are removed. Afterward, the most important features to calculate the EPC are retrieved. Finally, from the 800,000 EPCs and a total of 88 features for the whole country, 61 features and ≈ 740,000 entries were removed, ending up with a total of 25 features and ≈ 60,000 EPCs. Both continuous and discrete features are normalized into a [-1, 1] interval. The target prediction values are illustrated in Table 4 by documenting their maximum, minimum, mean, and standard deviation.

The ML models prediction targets are the annual heating energy needs (*Nic*), annual cooling energy needs (*Nvc*), annual primary energy needs (*Ntc*) (Table 1), and *R* ratio (Table 2). The database with 60,000 EPCs, is split by 33% (1/3) to build a model validation test set and 67% (2/3) to train the models. Three different ensemble regression algorithms from the Sci-Kit-learn library are explored: ET; MLP with 3 nodes and 25, 50, and 25 neurons; and GB tested with default parameters. The models are trained with an increasing number of features selected according to their k-best scores [42] and for each target output with k-best = 10, 15, 20, and 25 features (Table 3).

RMSE and R^2 scores are logged as performance metrics for the models' regression outputs (*Nic, Nvc, Ntc, R*). Performance results for the three regression algorithms are compared and the best-performing algorithm is selected accordingly. The selected model is then subject to a k-fold cross-validation process [33] with k=6 folds. This process splits the testing subset into multiple (k) folds and yields less biased performance results.

2.2. Multi-Objective optimization

For the MOO problem, a Non-dominated Sorting Genetic Algorithm II (NSGAII) [43] was used, since it demonstrated its efficiency for similar class problems in previous research [44,45]. The algorithm runs with a random generator object, a tournament selection method, and a population size of 25, for a maximum of 250 iterations or convergence. The MOO problem decision variables correspond to the EPC's possible retrofit solutions.

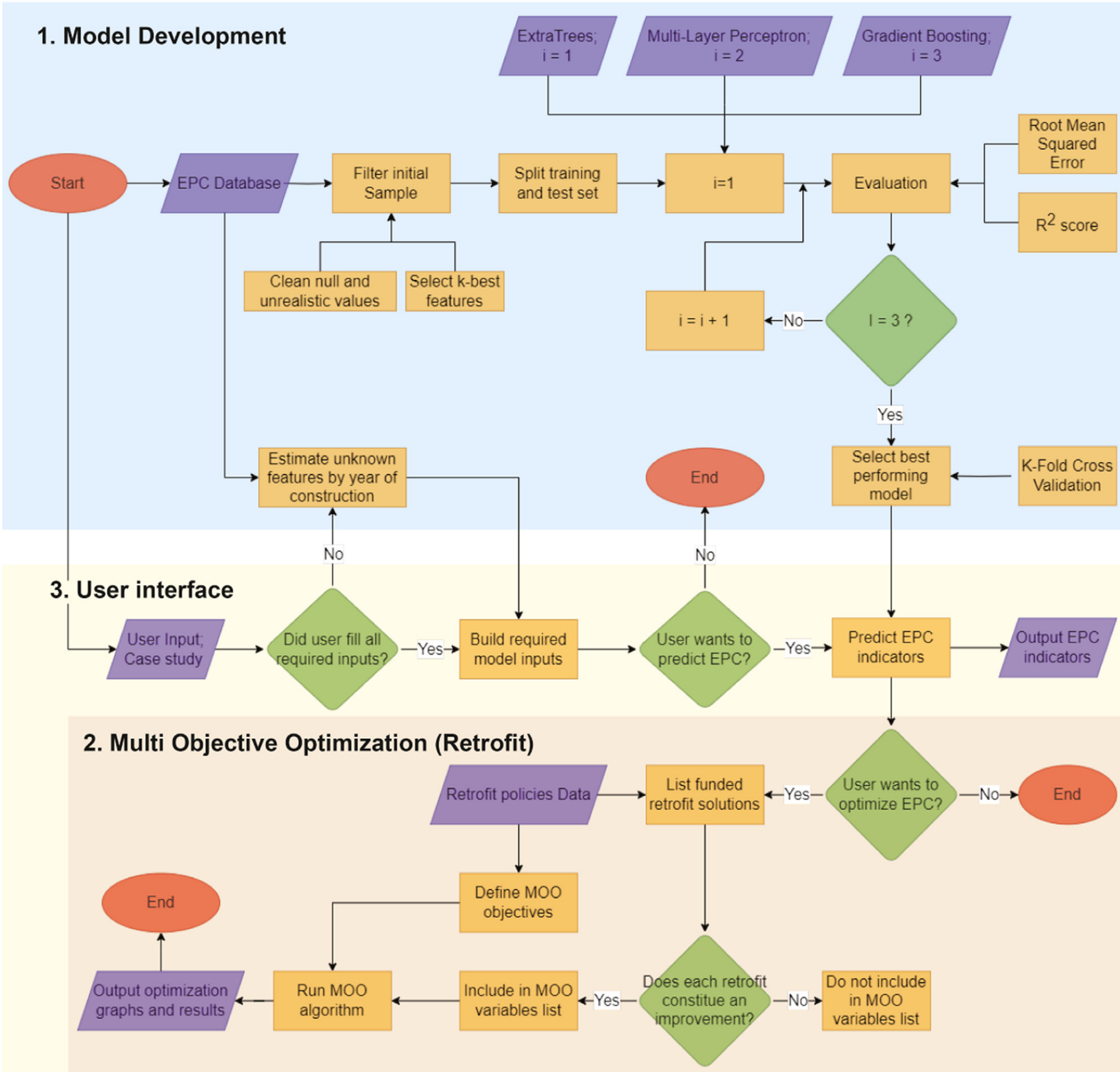


Fig. 1. Research workflow diagram.

The possible retrofits for one household are organized into different intervention areas: walls, roofs, windows, heating equipment, and Domestic Hot Water (DHW) equipment. A MOO process that returns the optimum retrofit options of an EPC is programmed, considering for each variable the option of having no retrofit and only those retrofits that represent an improvement against the original features. Finally, the total retrofit cost is calculated according to standard construction solutions, equipment costs, government funds, and typology (Tables 5 and 6, adapted from POSEUR [46]).

The optimization's objectives are the minimization of an EPC's *Ntc* (o_1 - Eq. 1) and retrofit cost (o_2 - Eq. 2), together with the maximization of the retrofit's Return on Investment over a period of 3 years (ROI) (o_3 - Eq. 3). o_1 is predicted by the surrogate model, o_2 is obtained from user inputs and retrofit costs from Table 5, and o_3 ROI is obtained from Eq. 4. The *Tax* variable represents the tax benefits obtained from the retrofit and is calculated by predicting the new EPC label and assigning the respective tax deduction. Energy Savings (E_s) is computed by Eq. 5, where p represents the average kWh price for electricity in Portugal (source: Eurostat [47]), A the floor area of the EPC, and t the number of years considered for the ROI calculation.

$$o_1 = min(Ntc_{new})\ [kWh/m^2] \tag{1}$$

$$o_2 = min(Cost)\ [€] \tag{2}$$

$$o_3 = min(ROI)\ [ratio] \tag{3}$$

$$ROI = \frac{E_s + Tax - Costs}{Costs}\ [ratio]; \tag{4}$$

$$E_s = pA(Ntc_{new} - Ntc_{old})t\ [€] \tag{5}$$

Table 5
Retrofit variables costs.

Retrofit Variable	Solution	Cost [€/m2]	Cost [€/unit]	Funding Ratio	Funding limit
Walls	Insulation	41	-	65	4500
Floors	Insulation	13.5	-	65	4500
Roof	Insulation (EPS)	13.5	-	65	4500
	Insulation (XPS)	25	-	65	4500
Glazing	PVC	260	-	85	1500
	Aluminium	380	-	85	1500
DHW	Gas boiler	-	450	0	0
	Water heater	-	175	0	0
	Boiler	-	1750	0	0
	Air-Water heat pump	-	3750	85	2500
Climatization	Gas boiler	-	450	0	0
	Water heater	-	175	0	0
	Boiler	-	2250	85	2500
	Multi-split	-	366	85	2500
Energy source	3 solar panels	-	6100	85	2500
	6 solar panels	-	9400	85	2500

Table 6
Multi-split equipment costs and government funds for each house typology.

Building typology	Costs [€]	Government funds ratio [%]	Government funds limit [€]
Studio	366		
1-bedroom	731		
2-bedroom	1096		
3-bedroom	1462	85	2500
4-bedroom	1828		
5-bedroom	2193		
6-bedroom	2558		
7-bedroom	2924		

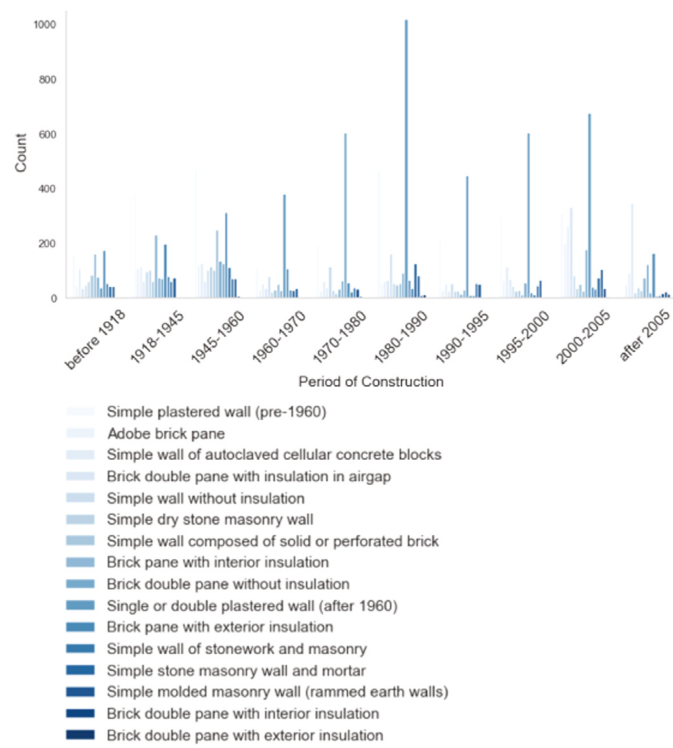


Fig. 2. Frequency of Wall type solutions for each construction period.

2.3. User interface - Case study

Initially, it is required that users fill in general information about the building/house regarding geometry, constructive solutions, heating, cooling, and DHW equipment (Fig. 1). The features belonging to Construction elements (Table 3) can be estimated based on the construction period of the household. Particularly, the Wall type, Roof type, and Floor type can be respectively estimated according to Figs. 2, 3, 4, and 5. With this data input, the model predicts the total, heating, and cooling annual energy needs, and EPC label. Users may eventually compare the predicted results with their billing information and original EPC, as well as test multiple retrofits manually.

After this initial exploration, an option to optimize the described EPC is available by setting a maximum budget and filling in their household's tax amount. The MOO algorithm starts to iterate over the possible retrofit scenarios according to the EPC data previously filled in by the user. Finally, results are illustrated in interactive bar charts for each objective, and a table with the optimum retrofit solutions for the maximum budget is presented. This interface was tested in a case study for a 3-bedroom household in Évora, Portugal (Fig. 6) with an EPC level of "F" (Fig. 7). This EPC was selected based on privacy concerns and because 3-bedroom households are the most common entry in the full database (≈ 45%). Most features are obtained from the house's technical drawings, while the unknown features are deduced according to its construction period. Feature values for the case study are documented in Table 7. Finally, the obtained optimization results are presented and discussed.

3. Results and discussion

This section is structured into four sub-sections: surrogate models' performance; MOO results; Case study results; and Discussion and Analysis. The first sub-section describes the performance of each tested surrogate model for the described k-best features. The models' performance is showcased by presenting their test sample's RMSE and R^2 for each target output. The MOO results sub-section displays the optimization results for one set of optimum retrofit options by mapping its objectives in a Pareto optimality chart [29]. The case study sub-section illustrates the front-end visualization and user interaction of both the surrogate model and optimization for the defined case study. Finally, the Discussion and Analysis section interprets the results and extrapolates the key advantages and limitations of this work by relating it to previous research.

3.1. Surrogate models' performance

The results obtained show that all algorithms perform similarly (Table 8). However, as the ET algorithm with 20 features tends to perform slightly better, it was selected for the final interface and MOO process. Specifically for the EPC prediction, the model developed with the ET algorithm has errors from -1 to +1 EPC level (Table 2, Section 1.1).

Table 7
Feature values (Table 3) for the proposed case study.

Details		Construction Elements		Equipment		Glazing	
Property	Apartment	**Wall type**	By Year	**DHW source**	Electricity	**Window area** $[m^2]$	9
Year	1996–2000	**Wall area** $[m^2]$	56	**Heating source**	Electricity	**Window type**	By Year
Area $[m^2]$	100	**Roof type**	By Year	**DHW type**	Heater		
Height $[m]$	3	**Roof area** $[m^2]$	116	**Heating type**	Split		
Typology	3 bedroom	**Floor type**	By year	**N°DHW equip**	1		
N°floors	2			**N°heating equip**	2		
District	Évora						

Table 8
Model training and performance indicators results.

		R [ratio]		Ntc [kWh/m²]		Nic [kWh/m²]		Nvc [kWh/m²]	
k-best features	Model	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
10	ET	0.74	0.26	0.58	55.70	0.61	19.11	0.31	5.85
	MLP	0.71	0.27	0.51	60.22	0.55	20.46	0.16	6.43
	GB	0.69	0.28	0.50	60.98	0.53	20.89	0.14	6.51
15	ET	0.82	0.22	0.72	45.30	0.65	18.07	0.36	5.63
	MLP	0.77	0.24	0.66	49.91	0.57	20.05	0.19	6.33
	GB	0.76	0.25	0.65	51.07	0.57	20.01	0.17	6.39
20	ET	0.84	0.21	0.79	39.77	0.67	17.70	0.41	5.41
	MLP	0.78	0.24	0.72	45.29	0.58	19.77	0.23	6.15
	GB	0.78	0.24	0.73	44.87	0.57	19.96	0.24	6.14
25	ET	0.85	0.20	0.80	38.83	0.73	15.96	0.61	4.40
	MLP	0.80	0.23	0.74	44.33	0.67	17.66	0.47	5.11
	GB	0.80	0.23	0.75	43.19	0.68	17.22	0.33	5.75

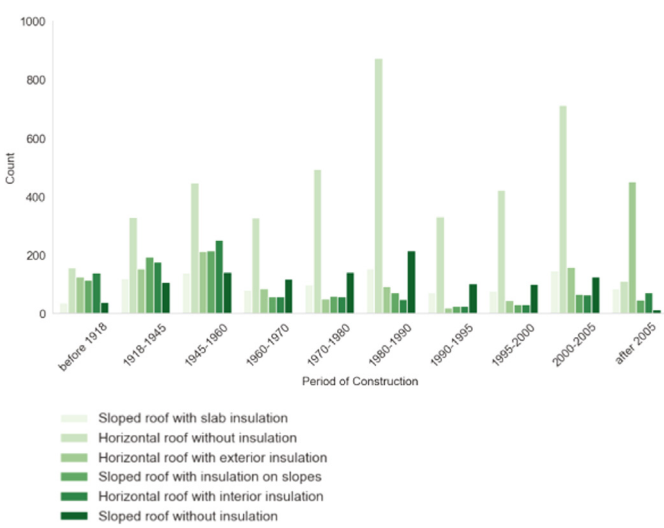


Fig. 3. Frequency of Roof type solutions for each construction period.

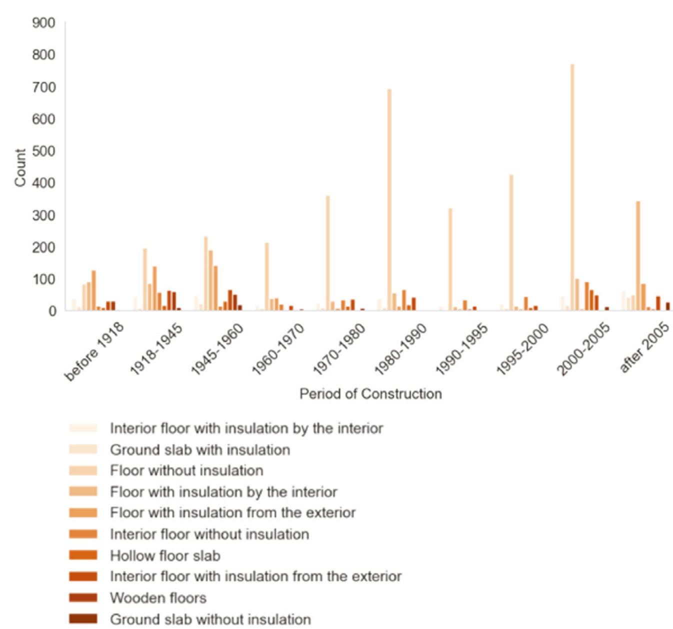


Fig. 4. Frequency of Floor type solutions for each construction period.

From the analysis of Table 8, it is possible to see that the R model and the Ntc model are the most accurate with an R^2 of 0.84 and 0.79, respectively. While Nic and Nvc show R^2 values of 0.67 and 0.41.

The selected algorithm was subject to a k-fold cross-validation process in which the full database was split into six training and test sets. Results from the selected algorithm are described in Table 9. R^2 values for Fold 1 and 2 are relatively lower than the remaining folds. The mean R^2 shows a slight decrease to 0.80 while the RMSE shows an increase to 0.24 when compared to the performance indicators of the selected ET algorithm with k-best = 20 features (Table 8). Nonetheless, the validation results still show acceptable accuracy, particularly for the R, Ntc, and Nic models that show R^2 scores ≥ 0.70.

Fig. 8 illustrates the error (in %) distribution plot of the selected algorithm for the prediction of the test samples (x-axis) and distribution of samples in the test subset for each target feature (y-axis). For the R prediction model, it is visible that most test values are distributed between 0 and 1, and respectively, the error values of the test set predictions vary between -25 and 25%. The Ntc, Nic, and Nvc prediction models show most test values distributed between 0 and ≈200, ≈100, and ≈20 kWh/m^2 respectively, while most error values vary between -50 to 50%. Additionally, a trend of higher error percentages for smaller target values is visible, which shows that most predictions that have an error of up to double the original value occur more frequently in smaller values. This model's behavior is seen in different scales for all models with the R prediction model being the least noticeable, and the Nic model the most.

Table 9
Cross validation results for the selected model ET with k-best = 20.

	R [ratio]		Ntc [kWh/m²]		Nic [kWh/m²]		Nvc [kWh/m²]	
	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Fold 1	0.61	0.24	0.63	49.46	0.55	21.27	0.34	7.26
Fold 2	0.79	0.28	0.76	50.83	0.67	21.43	0.50	6.99
Fold 3	0.86	0.25	0.79	54.54	0.70	22.27	0.46	6.00
Fold 4	0.87	0.24	0.84	51.21	0.78	20.87	0.51	5.86
Fold 5	0.85	0.22	0.81	45.52	0.75	18.32	0.43	5.33
Fold 6	0.83	0.22	0.80	44.24	0.75	19.20	0.45	5.10
Mean	0.80	0.24	0.77	49.37	0.70	20.48	0.45	6.07

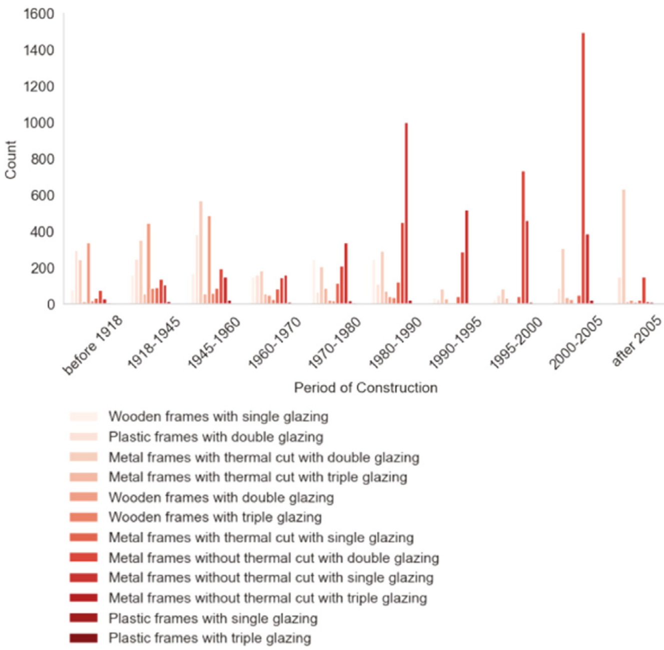


Fig. 5. Frequency of Window type solutions for each construction period.

Fig. 6. Case study 3D model.

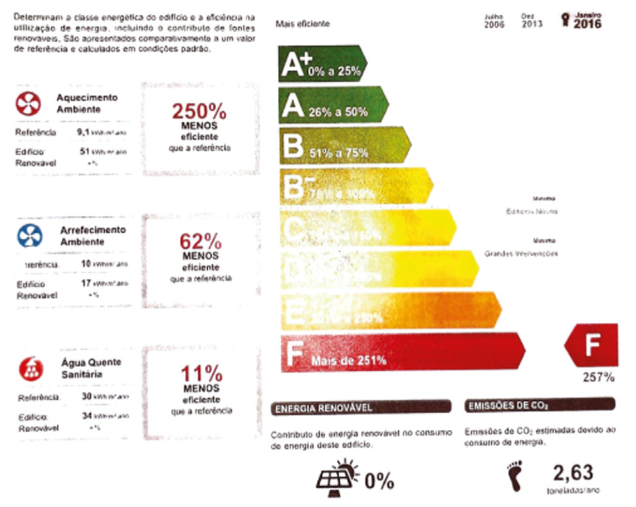


Fig. 7. Case study Energy Performance Certificate.

3.2. Multi-Objective optimization results

The optimization process was tested with the sample house described in Table 7 (Section 2.3) with an Ntc value of ≈ 22,500 kWh, a classification label of æEg, and no maximum budget. The results of the optimization process are described in a 3-dimensional Pareto optimality scatter chart represented in Fig. 9. The R ratio was added as a color scale of the optimal solutions to illustrate the relationships between objectives and their respective R. This chart illustrates the optimum solutions found by the algorithm, which are the non-dominated solutions (refer to Section 1.1 [29]). This chart is only presented here and is not included in the interface because this type of visual communication may be considered difficult to interpret by non-expert citizens.

Optimization results show a wide range of optimum solutions for all objectives. The objectives can vary between ≈11000 € and ≈450 € for the total retrofit costs, ≈6000 kWh and ≈17000 kWh for the Ntc, and ≈0.5 and 7 for the ROI in 3 years(i.e., return in 3 years of 0.5 to 7 times the investment made). This demonstrates that the best of each objective does not necessarily represent the best solution. In the case of this EPC, it is visible in Fig. 9 that the most expensive optimum solutions represent minimum Ntc and R values. However, solutions that have the highest ROI in 3 years correspond to the cheapest solutions with smaller R that have a more immediate impact on energy savings and tax benefits, when compared with the original features of the EPC. Finally, it is up to the user to choose the adequate ROI rate, how much should be spent on the retrofit, and if it is possible to plan long-term retrofit strategies based on returns and savings obtained from cheaper retrofits.

3.3. Case study interface results

The interface initiates by loading the developed surrogate models. After loading the models, drop-down and input boxes appear and allow the user to fill in basic general data regarding their home or building. The inputs for the proposed case study are filled in according to Table 7 (Section 2.3), with 16 mandatory features, and 4 optional, that can be extrapolated based on the construction period in the case of the information not being readily available (Wall, Roof, Floor, and Window types in Table 3). However, if the users can obtain the required details there is a possibility to fill in and change the default extrapolated values and increase the predictions' accuracy.

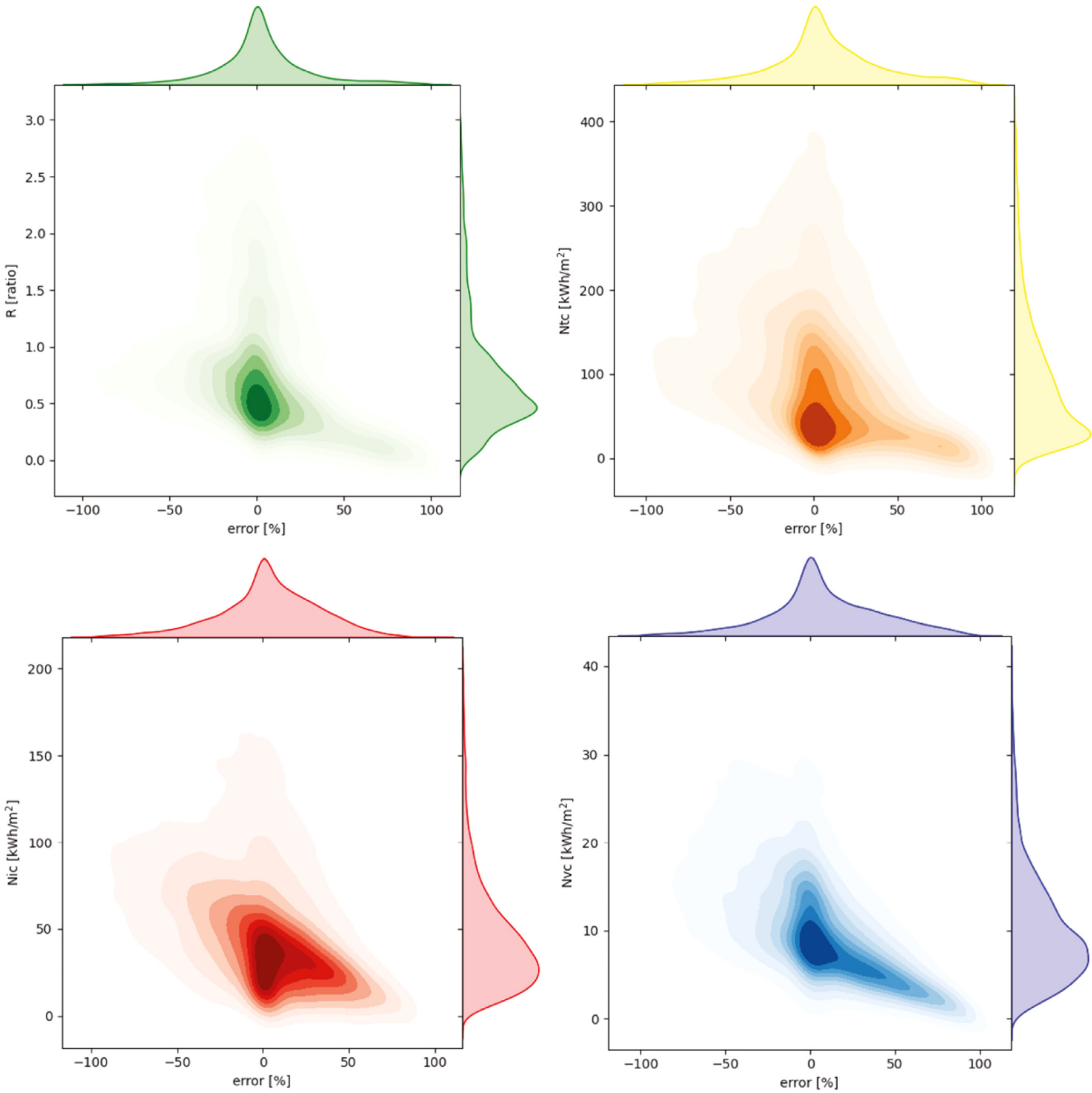


Fig. 8. Test set values (y-axis) and error (x-axis) distribution joint plot for each regression Extra Trees model.

Table 10
Case study predicted and original results.

	Nic [kWh/m^2]	Nvc [kWh/m^2]	Ntc [kWh/m^2]	R [ratio]	EPC class
Original	114.31	37.10	202.20	2.56	F
Predicted	102.92	40.99	225.05	2.10	E
error[%]	-10.49	9.96	10.69	-19.74	

After filling in the required data the user can simulate their energy needs Nic, Nvc, and Ntc values, and EPC (Fig. 10). The energy needs are presented in total kWh rather than kWh/m^2 for a better understanding and comparison with energy bills. Afterward, it is possible to manually change the inputs and perform analyses of retrofit impacts or keep scrolling to read the information regarding eventual government funds to support building retrofits, and tax benefits obtained from the EPC improvement.

The predicted results are compared with the original EPC for this house in Table 10. As seen, the predicted annual energy needs for heating (Nic), cooling (Nvc), and total (Ntc) have an error of $\approx 14\ kWh/m^2$, $\approx 3\ kWh/m^2$, and $\approx 23\ kWh/m^2$ respectively. Additionally, the model fails to predict the original EPC by one level. These errors are within the documented errors for the selected surrogate model in Table 8.

After this initial analysis, the user is prompted to click a button that performs an optimization process to find optimum retrofit solutions with

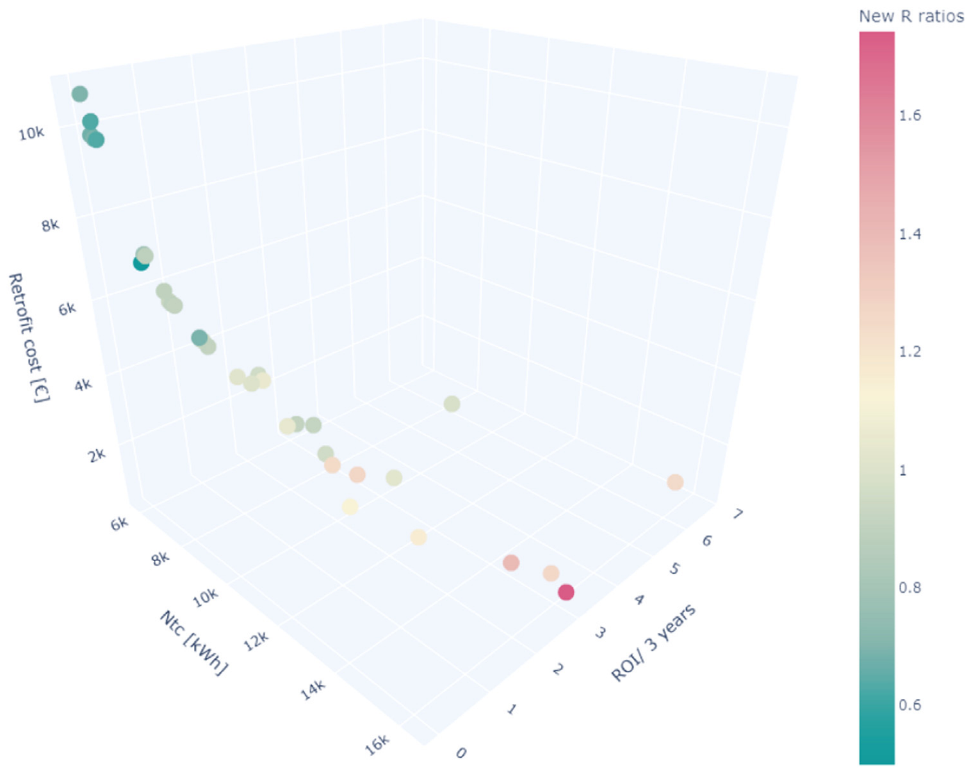


Fig. 9. Pareto optimality chart with a color scale of resulting retrofit R ratio.

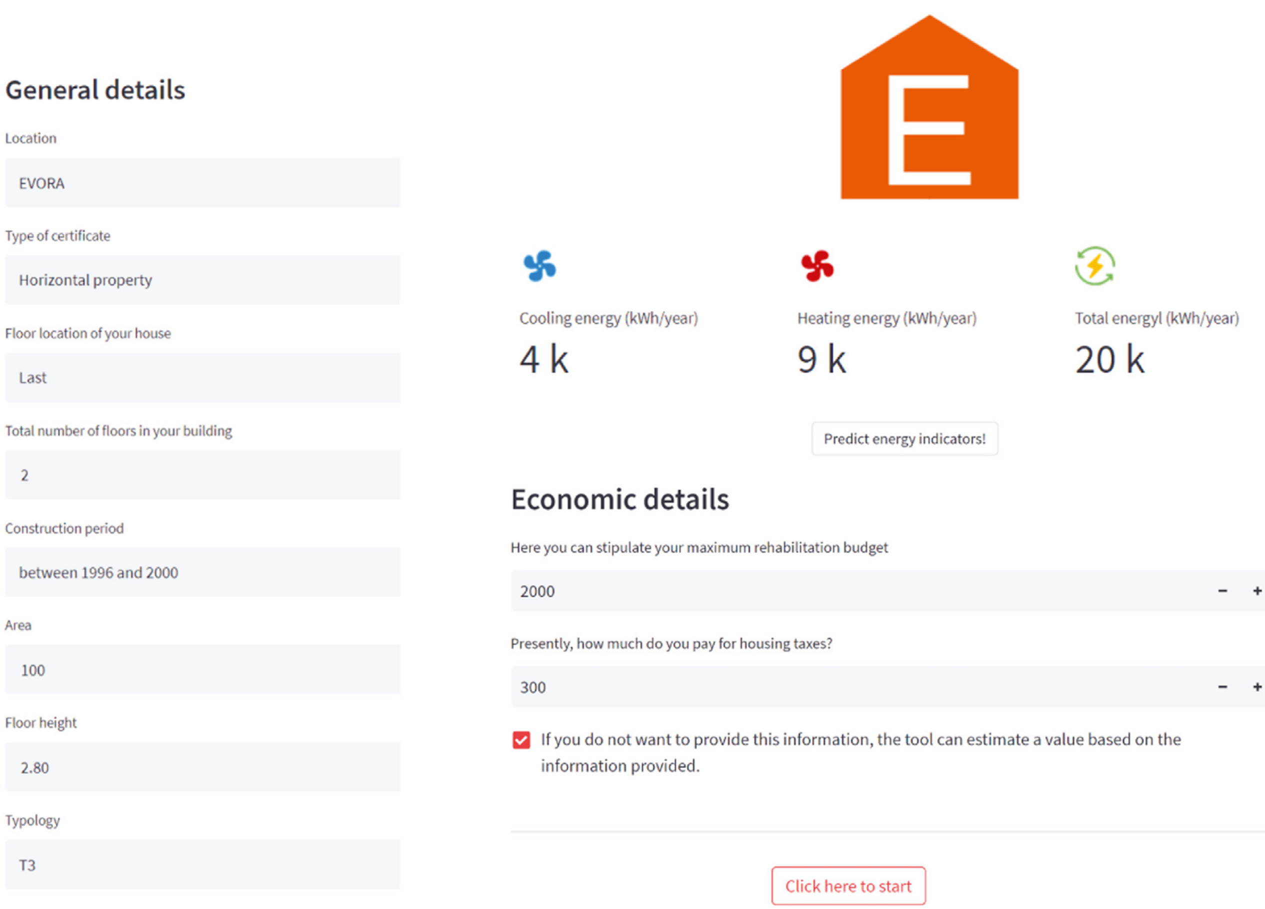


Fig. 10. Filler form for input features (left) and model predictions for EPC label (right).

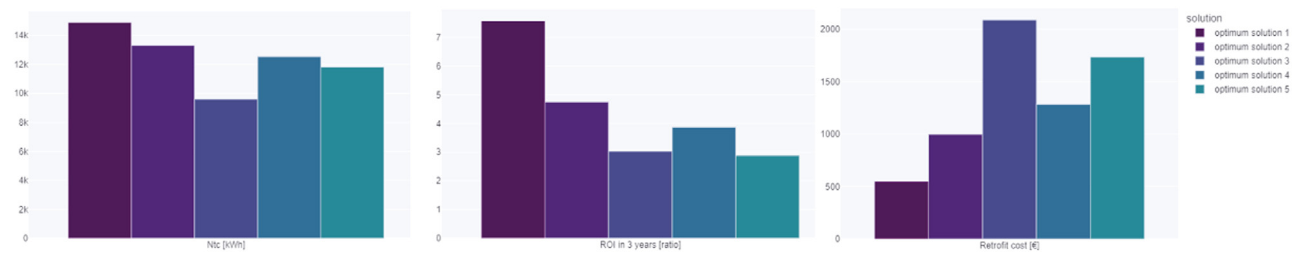


Fig. 11. Interactive charts for each objective: total energy needs (left). ROI in 3 years (middle), Retrofit cost (right). Table with optimum solutions values (Top).

Table 11
Example of the table with optimization results that can be downloaded by the user according to the specified budget limit.

Retrofit	Walls	Roof	Glazing	DHW	DHW energy source	HVAC	HVAC energy source	Ntc [kWh]	ROI [ratio]	Retrofit cost [€]	New EPC label
1	-	EPS	-	-	-	-	-	15473	7.32	548	C
2	-	EPS	-	Heater	Gas	-	-	13315	4.75	998	C
3	ETICS	EPS	PVC	-	-	-	-	9608	3.03	2088	B
4	-	EPS	PVC	-	-	-	-	12536	3.87	1284	C
5	-	EPS	PVC	Heater	Gas	-	-	11823	2.88	1734	C

a specific budget to be defined. In this case study, a maximum budget of 2000€ was selected, different from the optimization process performed above in sub-Section 3.2, which had no limit budget. The interface runs the optimization algorithm and provides the user with a table with the obtained optimum solutions that fit the specified budget. The user can interactively explore three bar charts that illustrate these solutions' objectives' values (Fig. 11). The bar charts were chosen as a more user-friendly option than the Pareto optimality chart to understand the results.

Results for this case study optimization show 5 non-dominated solutions that fit the specified budget (Table 11). Particularly, the algorithm suggested a retrofit of the roof, which currently has no insulation, followed by insulation of the exterior walls, which makes the retrofit closer to the budget limit. Cheaper variations of this solution are provided, such as changing from an electric water heater to a gas heater. The presented solutions can yield decreases in the Ntc by up to ≈ 60%, and ROI values ranging from ≈ 2 to ≈ 7.

3.4. Discussion and analysis

The results obtained in this work show R^2 scores of 0.84 and 0.79 and RMSE of 0.21 and 39.77 kWh/m^2 for the EPC (R ratio) and energy needs (Ntc) predictions respectively. These results are adequate for the purpose of integrating surrogate models and MOO into an easy-to-use web application that promotes the engagement of citizens and stakeholders in the building retrofit and EPC scheme. The validation results show slightly lower accuracy than the one seen in Burratti et al. [25] and Khayatian et al. [26] models. This may be explained by the databases used in these studies which may have different features and data. To improve the obtained accuracy in this study, new data cleaning methods can be used to improve additional features and feature engineering processes can be applied to develop additional features that might be missing from the original database.

The integration of the MOO process helped tackle conflicting objectives that have a significant weight in a citizen's decision about the rehabilitation of their household or building. For this case study EPC, the MOO process returned 5 optimum retrofit solutions capable of reducing energy savings up to 60% while minimizing the cost of the retrofit. However, optimization results should improve with the testing of multiple MOO algorithms, as well as with the tuning of each algorithm's hyperparameters. Additionally, an optimization process that explores different objectives and metrics may be helpful in the communication of results to the user.

Finally, the web app interface may improve with different communication strategies of both the input data, model exploration, and optimum retrofit solutions: (1) segmenting input and data forms into different sections; (2) assigning icons to inputs and features; (3) other interactive experiences such as interactive tables and/or charts. Furthermore, the web application would benefit from users' feedback, collaborative design, and integration of other EU countries' EPC databases. This would increase the scope and usability of the app to a larger scale.

4. Conclusions

This work integrates Machine Learning models and Multi-Objective Optimization into an interface in a web application format. This interface aims to engage citizens in the building retrofit process by promptly suggesting optimum retrofits for user-specific scenarios. In this sense, different regression algorithms were tested, and the best-performing ones were selected to predict energy needs and EPC labels with significantly fewer features than the ones required for the official certification process and for any building performance simulation or optimization. Additionally, this work provides researchers with alternative approaches to tackle energy-related problems, which are computationally inexpensive and have easy-to-grasp inputs.

The final models required 16 mandatory and 4 optional features to perform the predictions with a coefficient of determination of 0.84 and 0.79 for the EPC label and total energy needs predictions. Afterward, the optimization process considers all the retrofits that are funded by government programs and the user's maximum budget to find the best combinations of retrofit solutions that yield minimum energy needs and costs and maximum return on investment. Results have shown improvements of up to 60% decrease in energy needs and return on investments in 3 years of up to 7 for an EPC case study. Optimization results can then be consulted in the developed app as interactive tables and charts according to objectives and retrofit solutions. However, some limitations of this study lie in the accuracy of the surrogate model, the efficiency of the Multi-Objective Optimization results, and the interface data and results communication to the user.

In the future, multiple feature engineering techniques may improve the quality of the database and enhance the surrogate models' accuracy; Multiple optimization algorithms can be fine-tuned and compared to obtain the best optimization results; and new objectives and data communication techniques can be explored to further improve user experience. This work has the further advantage of allowing the replication of this approach to other national or municipal EPC databases. By applying the presented methodology, other countries or regions can develop better ways to provide citizens with valuable, accurate, and comprehensible information to enhance households' comfort and energy savings according to their investment budgets.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

G.R. Araújo: Writing – original draft, Methodology, Investigation, Conceptualization, Software. **Ricardo Gomes:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Paulo Ferrão:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **M. Glória Gomes:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization.

Acknowledgements

This work was supported by Fundação para a Ciência e Tecnologia (FCT) through IN+ UIDP/EEA/50009/2020 - IST-ID, through CERIS UIDB/04625/2020, and Ph.D. grant under the contract of FCT 2021.04849.BD. This research was funded by Project C-TECH - Climate Driven Technologies for Low Carbon Cities, grant number POCI-01-0247-FEDER-045919, LISBOA-01-0247-FEDER-045919, co-financed by the ERDF - European Regional Development Fund through the Operational Program for Competitiveness and Internationalization - COMPETE 2020, the Lisbon Portugal Regional Operational Program - LISBOA 2020 and by the FCT under MIT Portugal Program.

References

[1] J. Mourão, R. Gomes, L. Matias, S. Niza, Energy Build. 197 (2019) 34–46, doi:10.1016/j.enbuild.2019.05.033.
[2] O. Pasichnyi, J. Wallin, F. Levihn, H. Shahrokni, O. Kordas, Energy Policy 127 (2019) 486–499, doi:10.1016/j.enpol.2018.11.051. November 2018
[3] L. Pérez-Lombard, J. Ortiz, R. González, I.R. Maestre, Energy Build. 41 (3) (2009) 272–278, doi:10.1016/j.enbuild.2008.10.004.
[4] E. Parlament, J. Oficial n° L 153 (2010) 13–35.
[5] E. Parliament, J. Oficial da União Europeia L156 (2018) 75–91.
[6] E. Parliament, Energy Performance of Buildings Directive - EPBD (2002/91/EC), 2002.
[7] T. Nikolaou, D. Kolokotsa, G. Stavrakakis, Adv. Build. Energy Res. 5 (1) (2011) 53–70, doi:10.1080/17512549.2011.582340.
[8] M.A. Brown, A. Soni, M.V. Lapsa, K. Southworth, M. Cox, Progr. Energy 2 (4) (2020) 042003.
[9] K.T. Gillingham, P. Huang, C. Buehler, J. Peccia, D.R. Gentner, Sci. Adv. 7 (34) (2021) eabg0947.
[10] C. Reinhart, Z. Berzolla, H.D. Jacoby, J.M. Reilly, C. Schuler, Available at SSRN 3832381 (2021).
[11] Y. Nidam, A. Irani, J. Bemis, C. Reinhart, Environ. Plan. B: Urban Anal. City Sci. (2023). 23998083231154576
[12] J.P. Gouveia, P. Palma, Environ. Res. Lett. 14 (9) (2019), doi:10.1088/1748-9326/ab3781.
[13] E.N. ISO, Int. Org. Standard.: Geneva, Switzerland (2017).
[14] E.C.f. Standardization, En iso 52003-1, epb indicators, requirements, ratings and certificates part 1: General aspects and application to the overall energy performance (2017), 2017.
[15] E.C.f. Standardization, En iso 52010-1 epb, external climatic conditions part 1: Conversion of climatic data for energy calculations (2017), 2017.
[16] E.C.f. Standardization, En iso 52016-1 epb, energy needs for heating and cooling, internal temperatures and sensible and latent heat loads part 1: Calculation procedures(2017), 2017.
[17] E.C.f. Standardization, En iso 52018-1, epb indicators for partial epb requirements related to thermal energy balance and fabric features part 1: Overview of options (2017), 2017.
[18] E. Comission, DIRECTIVE 2010/31/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 19 May 2010 on the energy performance of buildings, 2010.
[19] D. Van Dijk, J. Hogeling, E3S Web of Conferences 111 (May) (2019) 6–14, doi:10.1051/e3sconf/201911104047.
[20] E. Commission, Renovation Wave: doubling the renovation rate to cut emissions, boost recovery and reduce energy poverty, Technical Report, European Comission, 2020. https://ec.europa.eu/newsroom/growth/items/690437
[21] A. Ferrantelli, J. Belikov, E. Petlenkov, M. Thalfeldt, J. Kurnitski, IEEE Access 10 (2022) 45430–45443.
[22] C. Watts, M.F. Jentsch, P.A.B. James, Build. Serv. Eng. Res. Technol. 32 (4) (2008) 361–376, doi:10.1177/0143624411404486.
[23] A. Hardy, D. Glew, Energy Policy 129 (2019) 1168–1178, doi:10.1016/j.enpol.2019.03.022.
[24] L.M. López-González, L.M. López-Ochoa, J. Las-Heras-Casas, C. García-Lozano, Appl. Energy 178 (2016) (2016) 308–322, doi:10.1016/j.apenergy.2016.06.028.
[25] C. Buratti, M. Barbanera, D. Palladino, Appl. Energy 120 (2014) 125–132, doi:10.1016/j.apenergy.2014.01.053.
[26] F. Khayatian, L. Sarto, G. Dall'O', Energy Build. 125 (2016) 45–54, doi:10.1016/j.enbuild.2016.04.067.
[27] Y. Fan, X. Xia, Appl. Energy 228 (April) (2018) 2140–2152, doi:10.1016/j.apenergy.2018.07.043.
[28] J. Khazaii, in: Advanced Decision Making for HVAC Engineers, Springer, Cham, 2016, pp. 87–97, doi:10.1007/978-3-319-33328-1_10.
[29] Y. Censor, Appl. Math. Optim. 4 (1) (1977) 41–59, doi:10.1007/BF01442131.
[30] A.-t. Nguyen, S. Reiter, P. Rigo, Appl. Energy 113 (2014) 1043–1058, doi:10.1016/j.apenergy.2013.08.061.
[31] T.G. Reames, M.A. Reiner, M.B. Stacey, Appl. Energy 218 (2018) 95–103.
[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, D. Edouard, et al., J. Mach. Learn. Res. 12 (2019) 2825–2830.
[33] T. Fushiki, Stat. Comput. 21 (2011) 137–146.
[34] S. García, J. Luengo, F. Herrera, ACM Comput. Surv. 50 (6) (2017) 163–193, doi:10.1007/978-3-319-10247-4_.
[35] J.a. Mendes-Moreira, C. Soares, A.M. Jorge, J.F. De Sousa, et al., ACM Comput. Surv. 45 (1) (2012), doi:10.1145/2379776.2379786.
[36] J. Feng, S. Lu, J. Phys. Conf. Ser. 1237 (2) (2019), doi:10.1088/1742-6596/1237/2/022030.
[37] A. Pinkus, in: Acta Numerica, volume 8, Cambridge University Press, 1999, pp. 143–195, doi:10.1017/S0962492900002919.
[38] J.H. Friedman, Comput. Stat. Data Anal. 38 (4) (2002) 367–378, doi:10.1016/S0167-9473(01)00065-2.
[39] R. Alizadeh, J.K. Allen, F. Mistree, Res. Eng. Des. 31 (3) (2020) 275–298, doi:10.1007/s00163-020-00336-7.
[40] T. Østergård, R.L. Jensen, S.E. Maagaard, Appl. Energy 211 (October 2017) (2018) 89–103, doi:10.1016/j.apenergy.2017.10.102.
[41] P.R. Diary, Decree-Law n° 80/2006, 2006.
[42] S. Begum, D. Chakraborty, R. Sarkar, in: 2015 International Conference on Computational Intelligence and Communication Networks (CICN), IEEE, 2015, pp. 811–814.
[43] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, et al., IEEE Trans. Evol. Comput. 6 (2) (2002) 182–197, doi:10.1109/4235.996017.
[44] G.R. Araújo, H. Teixeira, M.G. Gomes, A.M. Rodrigues, Sol. Energy 249 (October 2022) (2023) 446–456, doi:10.1016/j.solener.2022.11.043.
[45] I. Pereira, A. Leitão, Proc. Int. Conf. Architect. Sci. Assoc. 2020-Novem (November 2020) (2020) 765–774.
[46] POSEUR, Programa Operacional de Sustentabilidade e Eficiência no Uso de Recursos, 2020, (https://poseur.portugal2020.pt/). [Online; accessed 15-January-2020].
[47] Eurostat, Electricity price statistics, 2020, (https://ec.europa.eu/eurostat/). [Online; accessed 15-January-2020].