Energy & Buildings 285 (2023) 112922

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

Towards an airtightness compliance tool based on machine learning models for naturally ventilated dwellings

Vitor E.M. Cardoso [a,b,*], M. Lurdes Simões [b], Nuno M.M. Ramos [b], Ricardo M.S.F. Almeida [b,c], Manuela Almeida [d], Luís Sanhudo [a], João N.D. Fernandes [e]

[a] BUILT CoLAB – Collaborative Laboratory for the Future Built Environment, Rua do Campo Alegre, 760, Porto 4150-003, Portugal
[b] CONSTRUCT-LFC, Department of Civil Engineering, Faculty of Engineering, University of Porto, Porto 4200-465, Portugal
[c] Polytechnic Institute of Viseu, School of Technology and Management, Department of Civil Engineering, Campus Politécnico de Repeses, Viseu 3504-510, Portugal
[d] ISISE, Department of Civil Engineering (DEC), School of Engineering, University of Minho, Campus de Azurém, Guimarães 4800-058, Portugal
[e] FEUP, Faculty of Engineering, University of Porto, Porto 4200-465, Portugal

ARTICLE INFO

Article history:
Received 1 November 2022
Revised 21 December 2022
Accepted 17 February 2023
Available online 20 February 2023

Keywords:
Air change rate
Airtightness
Building energy conservation
Machine-learning
Multiple regression
Classifier ensemble

ABSTRACT

Physical models and probabilistic applications often guide the study and characterization of natural phenomena in engineering. Such is the case of the study of air change rates (ACHs) in buildings for their complex mechanisms and high variability. It is not uncommon for the referred applications to be costly and impractical in both time and computation, resulting in the use of simplified methodologies and setups. The incorporation of airtightness limits to quantify adequate ACHs in national transpositions of the Energy Performance Building Directive (EPBD) exemplifies the issue. This research presents a roadmap for developing an alternative instrument, a compliance tool built with a Machine Learning (ML) framework, that overcomes some simplification issues regarding policy implementation while fulfilling practitioners' needs and general societal use. It relies on dwellings' terrain, geometric and airtightness characteristics, and meteorological data. Results from previous work on a region with a mild heating season in southern Europe apply in training and testing the proposed tool. The tool outputs numerical information on the air change rates performance of the building envelope, and a label, accordingly. On the test set, the best regressor showed mean absolute errors (MAE) below 1.02% for all the response variables, while the best classifier presented an average accuracy of 97.32%. These results are promising for the generalization of this methodology, with potential for application at regional, national, and European Union levels. The developed tool could be a complementary asset to energy certification programmes of either public or private initiatives.

1. Introduction

1.1. Airtightness in the EPBD

An Energy Performance Building Directive (EPBD) energy certificate labels a dwelling according to its performance against a reference baseline [1]. Regarding air change rates (ACHs), the EPBD national transpositions often address the issue by imposing airtightness requirements.

Therefore, the chosen approach in the EPBD relates to the impact of the airtightness level on the variability of ACHs over time, as in leakier building envelopes, higher and less stable airflow volumes tend to occur, especially in naturally ventilated dwellings [2345].

Airtightness measures the resistance to inward or outward air leakage through unintentional leakage points in the building envelope [6]. In most of the naturally ventilated residential building stock the unintentional background leakages contribute to most, if not the entirety, of the ACHs [7]. These are often referred to as infiltration, and since they are unintentional, the quantification of their contribution is challenging.

Thus, when addressed, airtightness requirements are commonly bound to compliance with a limit validated by a blower door test or a prescriptive path during construction [8]. While the first is time and labour-intensive, the second is heavily reliant on assumptions. When not addressed, the EPBD national transpositions range from providing recommendations to experiencing a full omission on whole building airtightness requirements.

* Corresponding author at: BUILT CoLAB – Collaborative Laboratory for the Future Built Environment, Rua do Campo Alegre, 760, Porto 4150-003, Portugal.
E-mail address: v.cardoso@fe.up.pt (V.E.M. Cardoso).

https://doi.org/10.1016/j.enbuild.2023.112922
0378-7788/

One can explain this trend by the implicit clash of simultaneously maintaining the air change rates (ACHs) over a minimum level most of the time, for health and comfort reasons, while avoiding frequent high ACHs that potentially jeopardize energy efficiency [910]. Which, when accruing to the assumptions and efforts needed for addressment, helps to understand the lack of requirements in the latter.

1.2. ML in ACH research

To tackle the inherent conflict and lack of awareness in regions with predominantly mild climates, previous research [11] found airtightness performance ranges that effectively provide adequate ACHs in naturally ventilated dwellings through a labelling strategy.

From the required acquisition of the input data to the execution of airflow simulations to the labelling process, the developed research anchored itself in a complex and compound workflow, which does not favour widespread implementation nor ease its use by practitioners. What if the whole process could be simplified, mathematically modelled, and structured in a way that provides a complete, user-friendly tool that outputs information with low errors and high accuracy compared with the original setup?

In the last decade, machine learning (ML) models have gained greater application in building energy efficiency topics. These predictive models are natural add-ons to stochastic approaches [1213]. Among other applications, ML models reduce simulation time costs and often provide user-friendly tools to characterize an existing dwelling or design one [1415].

A review on the perspectives for the future of natural ventilation of dwellings indicates that surrogate models allow classifying dwellings per the ventilation performance, emphasizing the added value of combining passive strategies with renewable and sustainable energy solutions [16].

The literature presents several recent examples of models developed to predict the airtightness performance of building envelopes, based on Multiple Linear Regression (MLR) [17], clustering after Principal Component Analysis (PCA) [18], and Generalized Linear Models (GLM) [19], achieving moderate to considerable success.

Research on the effect of building characteristics on heating and cooling loads showed that random forests outperformed linear regressions in predicting their relationship [20]. The authors highlight the implicit advantages of decision tree mechanisms in dealing with statistical assumptions, such as multicollinearity between input variables [21].

A probabilistic approach to weather data and dwelling characteristics in the United Kingdom's built stock found artificial neural networks to be the best predictors of indoor air quality parameters [22]. The probabilistic approach overpasses several limitations of learning algorithms, such as training ranges and limited datasets [23].

Recently, an optimization system for building envelope design based on gradient boosting machines was successfully applied, providing mean absolute errors below 3.5 % between predicted and actual operative energy consumption [24]. Work on relating air infiltration predictions by converting ACH between 50 Pa and 4 Pa of pressure difference, based on the Persily-Kronvall model, only found modest coefficients of determination between actual and predicted values, despite the application of nonlinear fitting methods [25]. A regression tree designed from several Computational Fluid Dynamics (CFD) simulated values of ACH presented itself as a better substitute for the latter [26].

Gradient boosting models displayed higher accuracy than multivariate linear regression in predicting building ventilation potential in urban locations [27]. Random Forests (RFs) and Support Vector Machines (SVM) are some of the other applied surrogate models with the highest performance in predicting ventilation and energy performance [282930].

A Support Vector Regressor (SVR) requiring 29 building parameter inputs predicted the performance of approximately 50.000 detailed energy and airflow simulations with high accuracy [31]. Other implementations of SVR in building performance showed similar results [3233]. The authors highlight that the developed model could form a comfort performance labelling program for naturally ventilated commercial buildings.

1.3. Gaps and objectives

The connection between labelling programs and ML seems encouraging. When dealing with passive strategies, particularly natural ventilation, a single classification or regression output often does not compute a comprehensive characterization [31]. Still, a thorough literature review could only find one study [22] addressing infiltration and airtightness characteristics with detail to a broad built stock with ML techniques. The study refers to the United Kingdom, in line with the northern European context.

On a southern European mild climate context, the work from Cardoso et al. [11] outputted a large dataset on dwelling characteristics, their respective ACHs time series and labels. Thus, it provides the needed data for this research.

By developing and applying a methodology based on ML models to assess the air change rates performance of naturally ventilated dwellings, the main objective of the current research is to explore the potential of a tool in checking the airtightness compliance of these dwellings regarding health and energy efficiency issues. The present work aims to:

- Emulate the whole process of airflow balance for a representative meteorological dataset, reducing computational costs and time;
- Train regression models to predict a dwelling average ACH and the percentage of time the ACH is below, between, and above the lower (LL) and upper (UL) limits, respectively;
- Train classification models to predict the label of new dwellings, either non-compliant by default (NCd), compliant (Com), or non-compliant by excess (NCe);
- Provide the structure for a possible user-friendly airtightness compliance tool, which is the set of regression and classification trained models.

Since the initial dataset is the product of a simulation campaign, one expects the applied ML models to generally output low errors regarding regression and high accuracies regarding classification. However, using an air mass flow balance model for design or retrofit strategies is time and labour-intensive and requires expert knowledge. As such, it may not be the best approach in fulfilling practitioners' needs and general societal use, particularly regarding large-scale applications, i.e., regional or countrywide. Thus, the focus is on overcoming these limitations.

2. Methodology

The methodology presented in Fig. 1 maps the workflow rationale, the data used, and the processes followed to achieve the compliance tool. Two parts divide it: (1) dataset creation; and (2) machine learning framework. Subsections 2.1 and 2.2 describe these parts in detail, while subsection 2.3 adds information to the applied machine learning models.

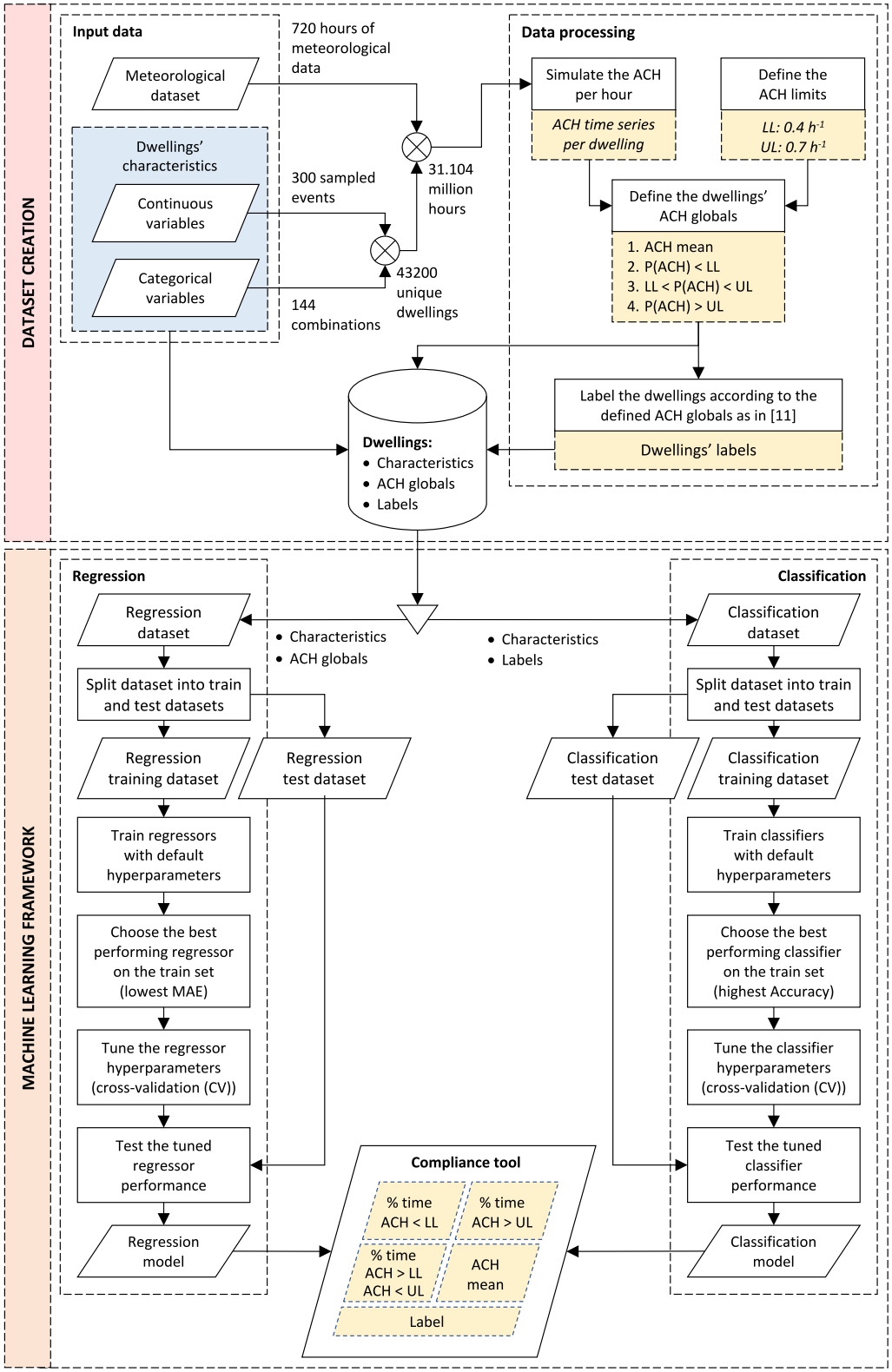

Fig. 1. Roadmap on the compliance tool, including the creation setup of the used dataset and the ML framework. P(ACH condition) stands for the percentage of time the ACH complies with the stated condition.

2.1. Creation of the used dataset

The input dataset encompasses the terrain and dwelling characteristics, the resulting ACH time series global descriptors, and the respective dwellings' airtightness performance labels that resulted from the probabilistic approach developed in [11]. The following describes the creation of this dataset, providing contextualization of the respective rationale behind it.

Table 1 presents the categorical and continuous input variables. These relate to geometry, terrain, and airtightness variables con-

Table 1
Geometry, terrain, and airtightness variables considered.

Categorical variables	Levels		
Location (terrain) (α)	0.14/0.22		
Side ratio (SR)	1:1/2:1		
Roof slope (RS)	0°/20°		
Number of exposed vertical surfaces (ES)	2/3/4		
Number of vertical ducts (VD)	2/3/4		
Number of floors (NF)	1/2		
Total of combinations	144		
Continuous variables	Distribution	Median	Std. Dev.
Floor Area (AF) [m^2]	LogNormal	129.24	70.13
Ceiling Height (CH) [m]	LogNormal	2.62	0.28
Airtightness at 50 Pa (n_{50}) [h^{-1}]	LogNormal	6.45	3.50
Airflow exponent (n) [–]	LogNormal	0.60	0.04

sidered to characterize the dwellings. The categorical variables agree with representative reference dwellings of the Portuguese built stock [34] to achieve a higher diversity of terrain and dwelling characteristics, totalling 144 combinations.

Spanish INFILES project [35] provided data on the continuous variables. These fitted to statistically significant lognormal distributions, from which a smart sampling method outputted 300 pseudo-random events. Comparison with previous Portuguese studies showed that the dwellings have similar overall distributions on airtightness (n_{50}) [11]. The product of the 300 events with the 144 discrete combinations provided a full dataset for simulation of 43,200 dwellings.

From the Porto/Pedras Rubras weather station, near Sá Carneiro Airport, in the Porto region of Portugal, for a representative month (720 h) of its heating season, the hourly average air temperature, wind speed and direction were used as meteorological data. The original data provides a probabilistic approach on the Portuguese residential built stock considering dwellings, terrain, and meteorology influential variables on air change rates across building envelopes, providing the needed data complexity to the input dataset of the present research.

Inputting the data in a single-zone air mass flow balance model allowed simulating the hourly ACHs for each dwelling. The process presented for airflow convergence has the same architecture as method 1 in the current version, EN 16798-7:2017 [36], the iterative method of the superseded EN 15242:2007 [37]. The MATLAB Optimization Toolbox [38] provided access to the implementation of the gradient descent method that performed the iterative process [39].

There are theoretical and practical limitations of the commonly applied types of air mass flow balance models in capturing the complexity of the natural ventilation phenomenon. In fact, a theoretical model such as the used single zone model is a compromise between: an empirical model that often overfit the original dataset it originates from [4041]; and the more complex theoretical models such as multizone air mass flow balance with Computational Fluid Dynamics (CFD) coupling [4243]. The latter are increasingly sensitive to discretization [44].

Common sources of uncertainty in the application of these models relate to accuracy issues regarding data on building leakage distribution, internal and external geometry properties, and weather [4445]. Two examples can be pointed: time steps on meteorological data readings often result in loss of information since the variability in the meantime is ignored [446]; and the use of wind pressure coefficients from wind tunnel studies with limited scope requires additional extrapolation leading to systematic errors in the final calculations [4748]. Still, even though airflow models simplifications, either in the model architecture, either in the input data, populate all simulated scenarios, they are widely applied and accepted, providing useful information for decision-making.

Since the dataset has time-invariant variables (those related to dwelling geometry and terrain characteristics) and time-variant variables (the meteorological variables and the dwellings ACHs over time), the latter were abstracted to global descriptors by applying the defined ACH limits to the outputted ACH time series.

A total of four global descriptors characterized the outputted ACH time series, namely: (1) ACH mean; (2) ACH's percentage of time below the defined LL (0.4 h^{-1}); (3) ACH's percentage of time between the LL and the UL (0.7 h^{-1}); and (4) ACH's percentage of time above the UL. The LL considered was 0.4 h^{-1}. The UL was defined at 0.7 h^{-1}. These correspond to the default predefined ventilation airflow rates for residential buildings of categories IV and I, respectively, of EN 16798–1:2019 [49]. These categories relate to indoor environmental quality and occupants' expectations.

To label the dwellings as compliant (Com), non-compliant by default (NCd), or non-compliant by excess (NCe), constraints on ACH dataset global descriptors applied, particularly the percentage of time the ACHs were above, below, or between the defined LL and UL (Fig. 2).

Those with an ACH below 0.4 h^{-1}>20 % of the time are labelled as non-compliant by default (NCd). This percentage aligns with category II of EN 16798-1 standard on the expected rate of dissatisfied occupants based on CO_2 levels [49]. From 1000 ppm, around 20 % of users are expected to feel dissatisfied with indoor air quality [50].

After applying the NCd threshold, the criterion shifts to the percentage of time the ACH is above the 0.7 h^{-1} upper limit for the remaining dwellings. A step increment of 2.5 % in time above this upper limit increment until a sample size of 5 % of the initial dataset gets encompassed. This group of dwellings is labelled as compliant (Com), and they are the top performers. The remaining dwellings are labelled as non-compliant by excess (NCe).

Obtaining a functional labelling strategy required ordering the features by relative importance, which helped better understand the impacts of meteorological, geometrical, airtightness, and terrain features. Most importantly, identifying and removing low relative importance features reduced the dataset size, helping to establish clear cut-off rules. Finally, this strategy classified the dwellings, which allowed to materialize airtightness performance ranges.

2.2. ML framework

For a comprehensive tool, and since surrogate models benefit from large datasets for training, this work uses the whole dataset, 43,200 dwellings. As hard labelling confers a significant loss of information on the amount of time the dwelling performs, in parallel, multi-output regression models apply in predicting the percentage of time the ACH of each dwelling is below 0.4 h^{-1}, between 0.4 h^{-1} and 0.7 h^{-1}, and above 0.7 h^{-1} and its respective mean ACH.

The dataset undergoes pre-processing operations regarding feature scaling and oversampling, which subsection 3.1 discloses. After, for both classification and regression, it divides randomly into training and test sets in a proportion of 80 % and 20 %, respectively, following the Pareto principle [51], a common practice in machine learning studies [5253], and theoretically supported [54].

In regression, a five K fold cross-validation (CV) applies to the training set for each model to assess the average Mean Absolute Error (MAE). A K fold cross-validation splits the dwellings into K equal-sized subsets, in other words, K consecutive folds, without shuffling. Each fold applies once as a validation set while the others form the training set. A K of 5 follows the same principle exposed in the last paragraphs since the data splits into an 80 % and 20 %

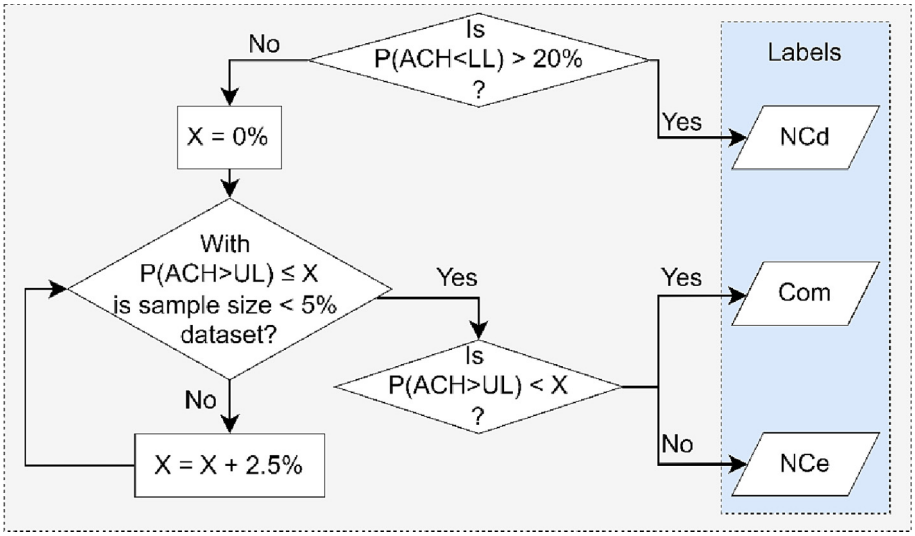

Fig. 2. Dwelling labelling strategy according to the defined lower (LL) and upper (UL) ACH limits.

proportion of the training set, in each fold. Additionally, a larger K introduces less bias in estimating the true expected error, but higher variance and a larger running time [5556]. Alternative CV procedures can be found in the literature [57].

In classification, a five stratified K-fold cross-validation is applied to the training set to determine the average accuracy. The stratified K-fold differentiates from the standard version by ensuring that the folds preserve the same percentage of events for each class, an important procedure to avoid bias in classification.

A grid search with cross-validation applies to the regression model with the lowest average MAE and the classification model with the highest average accuracy. A grid search uses all the possible combinations of the provided ranges of hyperparameters to fit and score a training set to a certain model. The method keeps the hyperparameters that output the best score. The resulting models with the tuned hyperparameters are used in both regression and classification to fit the test set.

In the proposed workflow (Fig. 1), the joint output of the best performing classifier, the one with the highest accuracy, and the best performing regressor, the one with the lowest MAE, will be a vector of five cells, with numerical information on the dwelling performance and a categorical label. The numerical output will be the three ACH percentages and the ACH mean. This approach fully characterizes the dwelling ACHs due to airtightness performance, including the outputted label's respective performance class.

2.3. Applied ML models

As reviewed in Subsection 1.2, both single and ensemble ML models provided promising results in replacing complete simulation setups. From the reviewed literature, SVM and DT-based methods obtained high performance in studies predicting ventilation, airflow, and energy performance [28293031], and as such these were the ones considered for the present research. Five models were applied, two single models: Support Vector Machines (SVM) and Decision Tree (DT); and three ensemble models: Random Forest (RF), Extreme Gradient Boosting (XGB) and Categorical Boosting (CB). These apply to both the classification and regression problems.

SVMs were first developed for classification purposes and rely on finding a suitable equation to divide an input dataset [58]. When facing two input features, the equation translates into a line, whereas, with more, the equation represents a hyperplane. The determination of the equation parameters corresponds to a convex optimization problem. Extension to regression changes the error function [59].

DTs encompass two main elements: branches and nodes [60]. Branches set true or false paths for the node conditions they depart. Nodes divide further into root, intermediate, and leaves. The root node uses the feature that best splits the data. The intermediate nodes can use the same feature as the root node or others [61]. Leaf nodes represent predictions, either a category in a classification setup or a numerical value in a regression one. After the learning process, the resulting flowchart is a roadmap for predicting newly presented observations. In classification, inspecting the impurity at each node informs on the quality of splits. Gini or entropy criteria usually apply in quantifying impurity. In regression, the mean squared error (MSE), or the mean absolute error (MAE) are the alternatives.

While a DT method offers increased interpretability compared to other single methods, such as SVMs, it has several disadvantages that result in a trend of the developed model overfitting the training data. At this point, ensemble methods are of interest, such as the case of RF [62]. A RF is an ensemble method that performs predictions by weighting several decision trees, commonly referred to as weak learners [63]. The weighting includes bootstrap resampling and random feature selection, among other methodologies. A randomly selected subsample from the training data trains each tree in bootstrap resampling. A reduced number of input variables are used at each node to split the observations in random feature selection. These processes mitigate the drawbacks of single DT models.

Proposed by Friedman [64], gradient boosted decision trees combine the advantages of RF, but instead of bagging homogeneous weak learners, they fit sequentially, adapting from the previous iterations. Each iteration focuses on the misclassified observations or those with the highest errors from the previous one. It can often lead to overfitting of the model to the training dataset. Furthermore, the sequential architecture increases computational costs compared to bagging methods. XGB [65] adds a regularisation term to the objective function, improving model generalization and reducing the number of iterations on finding the loss function minimum by computing the second partial derivatives.

Although CB [66] introduces changes in leaf growth, feature importance, and the encoding of categorical features, the main difference to XGB comes from the split criteria. CB applies Minimal Variance Sampling [67], which is a weighted version of stochastic gradient boosting [68] that resulted in increased quality models compared to XGB implementations [697071].

3. Results and discussion

3.1. Pre-processing

The correct implementation of the regression and classification models requires the execution of pre-processing operations. The procedures in this work were feature scaling and oversampling.

Regarding the first, a standard scaler centred the data at a mean of 0 and a standard deviation of 1. The data is then transformed according to the fitted scaler, ensuring that the original magnitude of each feature does not affect its impact on the model. The dwelling features in the test set transform according to the fitted scaler. The test set is not included in fitting the scaler, so it does not introduce bias in the training set.

Unbalanced label representation relates to the second procedure. Oversampling deals with the unbalanced groups that populate the dataset, which could result in several drawbacks when applying the classification models [72]. The groups of dwellings by label result from the described labelling limits and criteria. In a rural terrain, the numbers are 1257 dwellings labelled as Com, 10,377 as NCd, and 9966 NCe. In urban terrain, these numbers are 847 dwellings as Com, 13,482 as NCd, and 7271 as NCe. Descriptive statistics of the highest-performing dwellings are available in Table 2.

Thus, a Synthetic Minority Oversampling Technique for Nominal and Continuous features (SMOTE-NC) was applied, which allows for continuous and categorical features [73]. The process does not add information or variability to the original dataset. This solution overcomes the overfitting problem of random oversampling when classes have disparate representativeness [72]. As so, being NCd the majority class with 23,859 dwellings, the final input dataset is 71,577 dwellings, three times the majority class size, and equally divided by the three classes.

SMOTE-NC is a data augmentation technique that synthetically samples the minority classes into the size of the majority class based on nearest neighbours judged by the Euclidean distance between data points in feature space. It mainly differs from the default SMOTE technique by adding to the distance calculation the medians of the standard deviations of all continuous features for the minority class, if the nominal features of an instance are different to those of its potential nearest neighbours [73]. For datasets populated by continuous features only, there are several alternatives, such as BorderLine SMOTE [74], focused on synthesizing instances close to the target class boundaries, Adaptive Synthetic Sampling (ADASYN) [75], which synthesizes new instances according to the data density, among others [76]. More recently, the SMOTE-ENC [77] emerged as an alternative to treat mixed continuous and categorical datasets, from which one of the most relevant advantages compared to the SMOTE-NC is working correctly without any continuous feature, while the latter requires at least one.

Table 2
Descriptive statistics of the groups of highest performing dwellings, the compliant (Com) labelled, in a rural and an urban terrain for the whole dataset.

Feature	Metric	Rural (N = 1257)	Urban (N = 847)
ACH [h^{-1}]	**Mean**	**0.65**	**0.55**
	Std. dev.	**0.25**	**0.15**
ACH < 0.4 h^{-1}	**Mean**	**14.24 %**	**13.27 %**
	Std. dev.	**4.69 %**	**4.82 %**
0.4 h^{-1} < ACH < 0.7 h^{-1}	**Mean**	**52.08 %**	**72.44 %**
	Std. dev.	**6.43 %**	**5.40 %**
ACH > 0.7 h^{-1}	**Mean**	**33.68 %**	**14.29 %**
	Std. dev.	**5.25 %**	**3.92 %**
n_{50} [h^{-1}]	**Mean**	**5.36**	**5.86**
	Std. dev.	**1.21**	**1.35**
n [–]	Mean	0.60	0.60
	Std. dev.	0.04	0.04
AF [m^2]	Mean	124.79	101.06
	Std. dev.	57.30	32.98
CH [m]	Mean	2.62	2.61
	Std. dev.	0.21	0.26
SR [–]	Mean	1.45	1.50
RS [°]	Mean	10.64	12.04
ES [–]	Mean	3.15	3.12
NF [–]	Mean	1.61	1.54
ND [–]	Mean	3.19	3.33

3.2. Multi output regression

Table 3 presents the average training, CV mean, and CV standard deviation MAE from applying the considered regression models. Only test set predictions portray the separate MAE for each output. Therefore, since not all of them have the same unit, the MAEs presented in Table 3 also do not present one.

The single models all performed worse than the ensembles. Nonetheless, the DT performed marginally worse than the XGB regressor. RFR and CBR performed similarly, with CBR getting the lowest average CV MAE and showcasing less overfitting to the training data from the two. For this model, tuning occurs separately for the number of trees, the maximum depth, and the learning rate to identify the ranges for grid search (Fig. 3).

The number of trees identified was 2000 and 3000. The maximum depths identified were 10 and 12. The learning rates determined were 0.3 and 0.4. The grid search between the eight combinations, totalling 40 folds, resulted in the best parameters being 3000 trees, with a maximum depth of 10 and a learning rate of 0.3. The corresponding CV average MAE was 0.79 %, with a standard deviation of 0.007 %. It represents 50 % of the MAE before tuning the hyperparameters.

As a final evaluation, one looks into the MAEs of the final model in predicting: the ACH mean, the percentage of time below 0.4 h^{-1}, between 0.4 and 0.7 h^{-1}, and above 0.7 h^{-1}, with the test set. Table 4 presents the MAEs. Fig. 4 plots the predicted and actual outputs from the model application to the test set.

As the ACH between 0.4 and 0.7 h^{-1} has two boundaries compared to the other two with only one each, the MAE is higher in the first. Still, overall, tuning the best-performing model results in a robust regressor with the MAE never exceeding 1.02 % of the time the ACH of a dwelling is within a certain range. Regarding the ACH mean, the MAE is residual.

3.3. Multi output classification

Table 5 displays the average training, CV mean, and CV standard deviation accuracy from applying the considered classification models.

Table 3
Average training MAE, CV average MAE, and standard deviation MAE for the considered regression models.

MAE [–]	Average training MAE	CV average MAE	CV std. dev. MAE
SVR	0.0456	0.0466	0.0004
DTR	0.0000	0.0241	0.0005
XGBR	0.0177	0.0210	0.0003
RFR	0.0058	0.0173	0.0001
CBR	0.0140	0.0156	0.0002

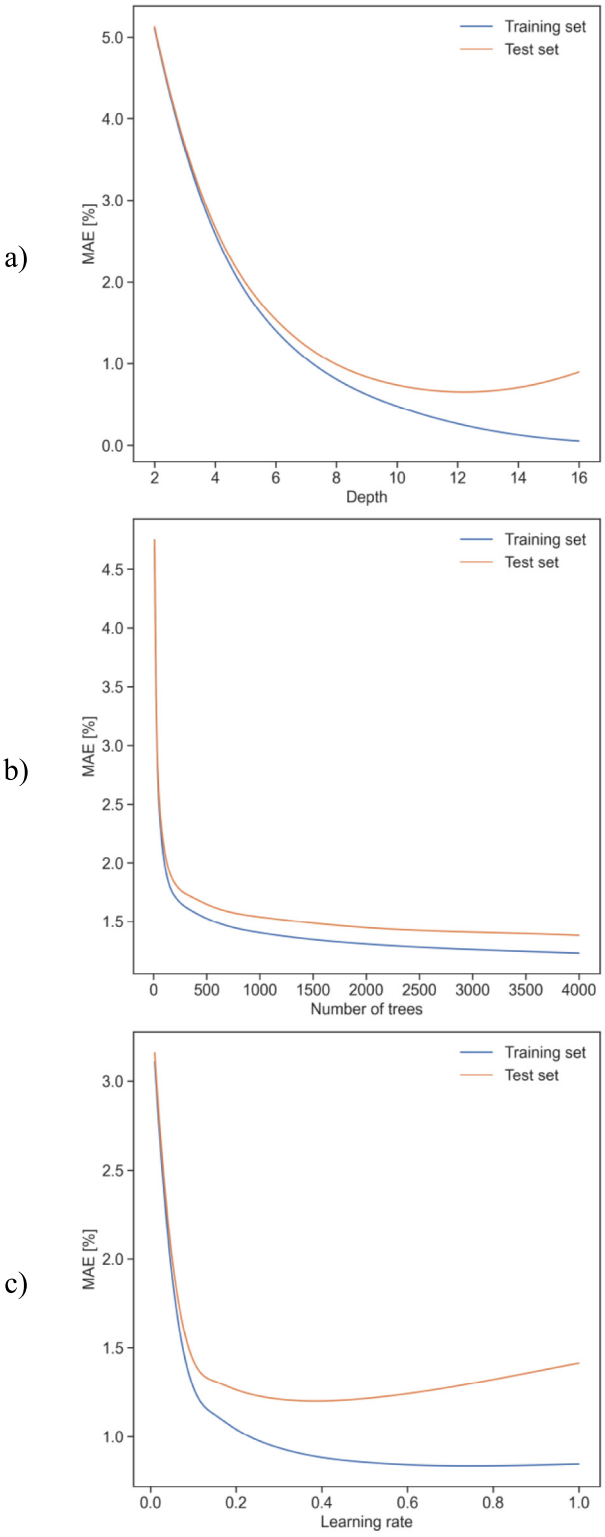

Fig. 3. Standalone CB regressor hyperparameters tuning using CV: a) the number of trees; b) maximum depth of trees; c) learning rate.

Again, the single models all performed worse than the ensembles. Nevertheless, DT performed marginally worse than the RF classifier. The ensemble methods performed similarly, with XGBC getting the highest average CV accuracy and the lowest standard deviation between fold scores. For this model, tuning of the number of trees, the maximum depth of each, and the learning rate occurred separately to identify the ranges for grid search (Fig. 5).

The number of trees identified was 150 and 200. The maximum depths determined were 7 and 8. The learning rates were 0.4 and 0.5. The results of the grid search in the eight combinations, totalling 40 folds, resulted in the best parameters being 150 trees, with a maximum depth of 7 and a learning rate of 0.4. This best estimator's average CV accuracy was 97.27 %, with a standard deviation of 0.08 %.

The final evaluation concerns its application to the test set. The classification metrics (Table 6) and the confusion matrix (Fig. 6) for these predictions portray the model performance on the test set.

Since classes NCd and NCe share the least similarity, the risk of misclassification is lower. More false positives are present in class Com, as precision scores are lower than in the other classes. Still, this class scores the highest recall, the ratio of correctly predicted

Table 4
Resulting MAEs from predicting the test set outputs with the best-performing hyper-tuned regression model. P(ACH condition) stands for the percentage of time the ACH complies with the stated condition.

Output	MAE
ACH mean	0.003 h^{-1}
P(ACH < 0.4 h^{-1})	0.82 %
P(0.4 h^{-1} < ACH < 0.7 h^{-1})	1.02 %
P(ACH > 0.7 h^{-1})	0.63 %

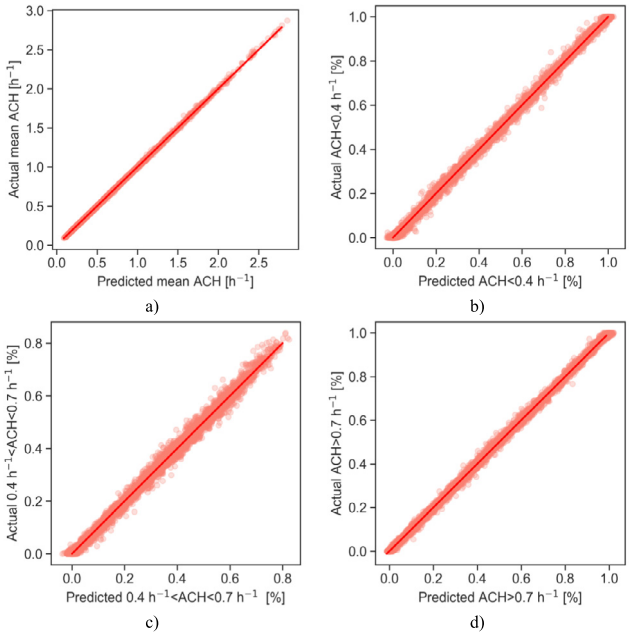

Fig. 4. Predicted versus actual outputs with the test set.

Table 5
Average training accuracy, CV average accuracy, and standard deviation accuracy for the considered classification models.

Model	Average training Accuracy	CV average accuracy	CV std. dev. accuracy
SVC	85.43	84.98	0.43
DTC	99.89	94.99	0.15
RFC	99.89	96.19	0.17
CBC	98.03	96.72	0.12
XGBC	98.60	96.78	0.18

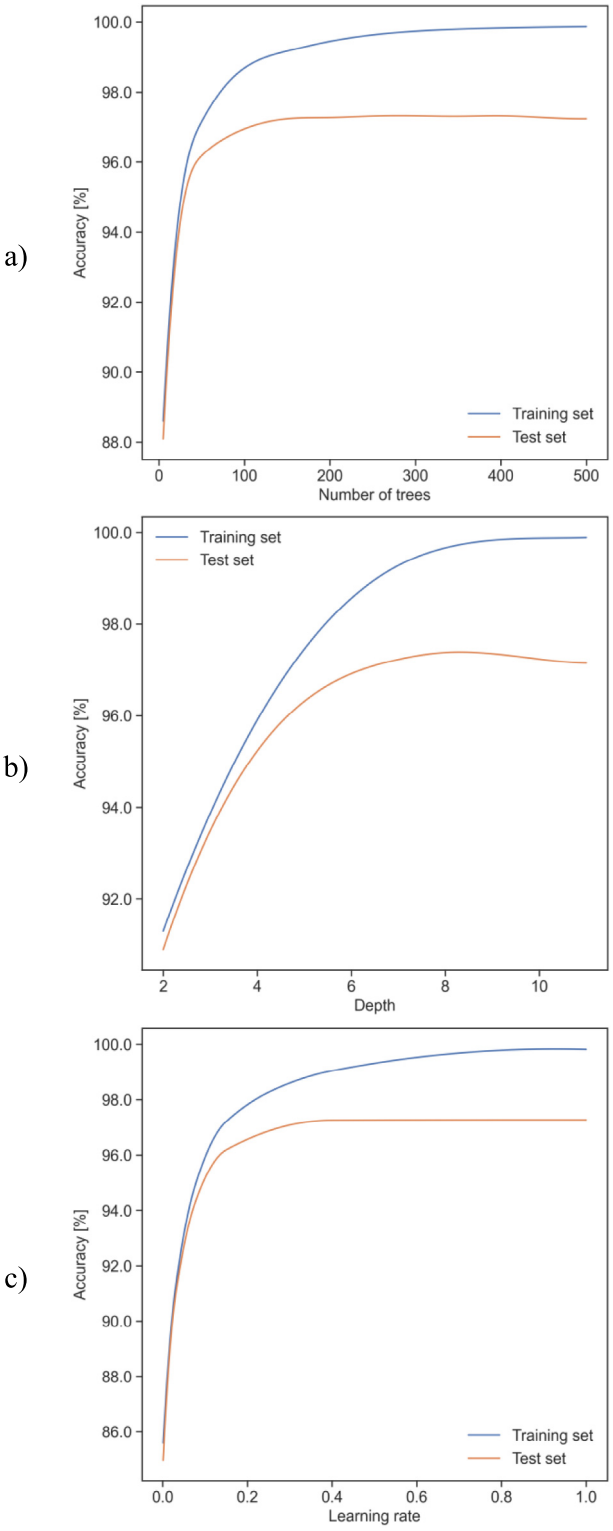

Fig. 5. Standalone XGB classifier hyperparameters tuning using CV: a) the number of trees; b) maximum depth of trees; c) learning rate.

observations to all the true observations in the same class. It is the preferable outcome since one prefers to label dwellings as compliant if they are close to the boundaries of compliance in a non-compliant group. The alternative is to label the dwelling as non-compliant when it is compliant. Overall, the achieved accuracy of 97.32 % in test set predictions supports the robustness of the classification model.

Table 6
Classification metrics resulting from the prediction of the test set classes with the best-performing hyper tuned classification model.

Labels	Precision	Recall	F-Score	Support
NCd	98.04 %	97.12 %	97.58 %	4695
Com	95.89 %	97.67 %	96.77 %	4725
NCe	98.06 %	97.18 %	97.62 %	4896
Average	97.33 %	97.33 %	97.32 %	14,316

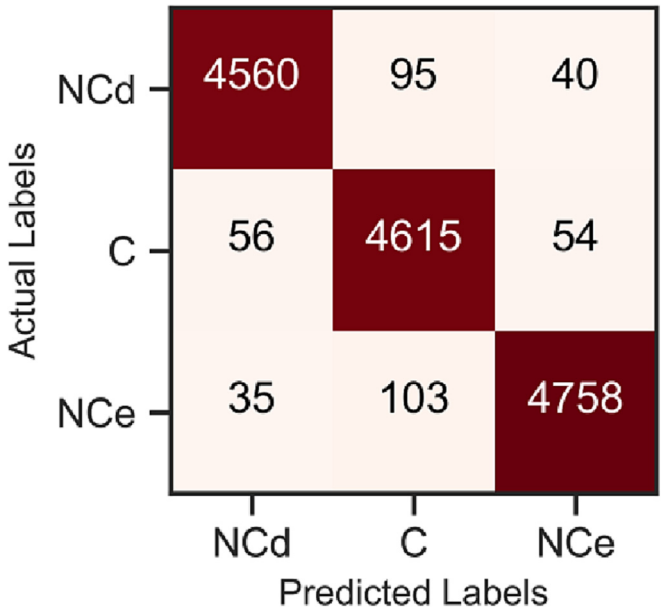

Fig. 6. Confusion matrix of the test set.

4. Conclusions

The current research presented a thorough workflow on the creation of a tool through a Machine Learning framework to check airtightness compliance in naturally ventilated dwellings and successfully validated it. For training models of such a tool, it needs a dataset on buildings characteristics, a meteorological dataset, and the time series from the resulting air exchanges. The present research used datasets of previous work as case study.

The tool communicates the air change rates performance and classifies a dwelling, in the studied geography, according to a labelling strategy. It requires a dwelling's geometric features and the result of an airtightness test to be inputted into the set of trained models to obtain an estimate of air change rates performance and a corresponding label. It successfully reduces the needed knowledge expertise, time, labour, and computational power needed to assess air change rates performance and airtightness compliance, without incurring in large errors and inaccuracies.

From the existing Machine Learning models identified as promising in the literature, this research applied a total of five. These models focused on the regression capabilities of predicting mean ACH, percentage of time below the defined lower ACH limit, between the lower and the upper limits, and above the upper ACH limit. Regarding classification, the models focused on labelling the dwellings as non-compliant by default (NCd), non-compliant by excess (NCe), and compliant (Com). The drawn conclusions are as follows:

- The developed methodology produced robust predictions. The regression model identified as the best predictor was the Categorical Boosting Regressor (CBR). The test set MAE on the per-

centages of time predicted with this model was: 0.82 % for the ACH below 0.4 h^{-1}; 1.02 % for the ACH between 0.4 and 0.7 h^{-1}; and 0.63 % for the ACH above 0.7 h^{-1}. Regarding the mean ACH, the MAE was 0.003 h^{-1};
- The classification model identified as the best predictor was the eXtreme Gradient Boosting Classifier (XGBC). The average accuracy of the test set, predicted with this model, was 97.33 %, with an F-score of 97.58 %, 96.77 %, and 97.62 % for dwellings non-compliant by default (NCd), compliant dwellings (Com), and dwellings non-compliant by excess (NCe), respectively.

The results of both regression and classification methods demonstrate robustness by having low errors and mispredictions. The used case study showcased the potential of applying the developed methodology, encouraging future works towards its application to different case studies for further validation.

Generalization requires an extensive campaign for wider meteorological time frames and dwellings datasets. In this work, the trained models relied on a single month representative of the heating season in a southern European mild climate. In the long run, the presented roadmap can fully map the meteorological conditions of several locations and typify the diversity of dwellings and terrain characteristics. It would form the groundwork for a possible airtightness compliance tool, with potential for application at regional, national, and EU levels, as a complementary asset to energy certification programs of either public or private initiatives.

Data availability

The data that has been used is confidential.

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.

Acknowledgements

This work was financially supported by: Base Funding – UIDB/04708/2020 and Programmatic Funding – UIDP/04708/2020 of the CONSTRUCT – Instituto de I&D em Estruturas e Construções – funded by national funds through the FCT/MCTES (PIDDAC). The author would like to acknowledge the support of FCT – Fundação para a Ciência e a Tecnologia, the funding of the Doctoral Grant PD/BD/135162/2017, through the Doctoral Programme EcoCoRe. This work is supported by the European Social Fund (ESF), through the North Portugal Regional Operational Programme (Norte 2020) [Funding Reference: NORTE-06-3559-FSE-000176].

References

[1] E. Comission, DIRECTIVE (EU) 2018/844 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 30 May 2018 amending Directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency, Off. J. Eur. Union 156 (2018) 75–91.
[2] A. Sfakianaki, K. Pavlou, M. Santamouris, I. Livada, M.-N. Assimakopoulos, P. Mantas, A. Christakopoulos, Air tightness measurements of residential houses in Athens, Greece, Build. Environ. 43 (4) (2008) 398–405.
[3] V.E.M. Cardoso, P.F. Pereira, N.M.M. Ramos, R.M.S.F. Almeida, The Impacts of Air Leakage Paths and Airtightness Levels on Air Change Rates, Buildings 10 (3) (2020) 55, https://doi.org/10.3390/buildings10030055.
[4] V.E.M. Cardoso, M.L. Simões, N.M.M. Ramos, R.M.S.F. Almeida, M. Almeida, L. Conceição, Impact of atmospherical stability and intra-hour variation of meteorological data in the variability of building air change rates, Build. Environ. 207 (2022), https://doi.org/10.1016/j.buildenv.2021.108528.
[5] R.M.S.F. Almeida, E. Barreira, P. Moreira, A Discussion Regarding the Measurement of Ventilation Rates Using Tracer Gas and Decay Technique, Infrastructures 5 (10) (2020) pp, https://doi.org/10.3390/infrastructures5100085.
[6] G. Guyot, R. Carrié, and P. Schild, "Stimulation of good building and ductwork airtightness through EPBD," *ASIEPI Intell. Energy Eur.*, no. March, 2010.
[7] M. W. Liddament. "A guide to energy efficient ventilation," *Air Infiltration Vent. Center, (AIVC).* 252. 1996.
[8] I. Poza-Casado, V.E.M. Cardoso, R.M.S.F. Almeida, A. Meiss, N.M.M. Ramos, M.Á. Padilla-Marcos, Residential buildings airtightness frameworks: A review on the main databases and setups in Europe and North America, Build. Environ. 183 (2020), https://doi.org/10.1016/j.buildenv.2020.107221.
[9] H.R.R. Santos, V.M.S. Leal, Energy vs. ventilation rate in buildings: A comprehensive scenario-based assessment in the European context, Energy Build. 54 (2012) 111–121, https://doi.org/10.1016/j.enbuild.2012.07.040.
[10] S. Nabinger, A. Persily, Impacts of airtightening retrofits on ventilation rates and energy consumption in a manufactured home, Energy Build. 43 (11) (2011) 3059–3067, https://doi.org/10.1016/j.enbuild.2011.07.027.
[11] V.E.M. Cardoso, M. Lurdes Simões, N.M.M. Ramos, R.M.S.F. Almeida, M. Almeida, J.N.D. Fernandes, A labelling strategy to define airtightness performance ranges of naturally ventilated dwellings: an application in southern Europe, Energy Build. (2022), https://doi.org/10.1016/j.enbuild.2022.112266.
[12] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renew. Sustain. Energy Rev. 81 (2018) 1192–1205, https://doi.org/10.1016/j.rser.2017.04.095.
[13] I. Jaffal, C. Inard, A metamodel for building energy performance, Energy Build. 151 (2017) 501–510, https://doi.org/10.1016/j.enbuild.2017.06.072.
[14] L. Van Gelder, P. Das, H. Janssen, S. Roels, Comparative study of metamodelling techniques in building energy simulation: Guidelines for practitioners, Simul. Model. Pract. Theory 49 (2014) 245–257, https://doi.org/10.1016/j.simpat.2014.10.004.
[15] T. Østergård, R.L. Jensen, S.E. Maagaard, A comparison of six metamodeling techniques applied to building performance simulations, Appl. Energy 211 (2018) 89–103, https://doi.org/10.1016/j.apenergy.2017.10.102.
[16] N.R.M. Sakiyama, J.C. Carlo, J. Frick, H. Garrecht, Perspectives of naturally ventilated buildings: A review, Renew. Sustain. Energy Rev. 130 (2020), https://doi.org/10.1016/j.rser.2020.109933.
[17] B. Khemet, R. Richman, An empirical approach to improving preconstruction airtightness estimates in light framed, detached homes in Canada, J. Build. Eng. 33 (2021), https://doi.org/10.1016/j.jobe.2020.101433.
[18] J. Feijó-Muñoz, C. Pardal, V. Echarri, J. Fernández-Agüera, R. Assiego de Larriva, M. Montesdeoca Calderín, I. Poza-Casado, M.Á. Padilla-Marcos, A. Meiss, Energy impact of the air infiltration in residential buildings in the Mediterranean area of Spain and the Canary islands, Energy Build. 188-189 (2019) 226–238.
[19] I. Poza-Casado, P. Rodríguez-del-Tío, M. Fernández-Temprano, M.-Á. Padilla-Marcos, A. Meiss, An envelope airtightness predictive model for residential buildings in Spain, Build. Environ. (2022), https://doi.org/10.1016/j.buildenv.2022.109435.
[20] A. Tsanas, A. Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, Energy Build. 49 (2012) 560–567, https://doi.org/10.1016/j.enbuild.2012.03.003.
[21] E. Tuv, A. Borisov, G. Runger, K. Torkkola, Feature selection with ensembles, artificial variables, and redundancy elimination, J. Mach. Learn. Res. 10 (2009) 1341–1366.
[22] P. Das, C. Shrubsole, B. Jones, I. Hamilton, Z. Chalabi, M. Davies, A. Mavrogianni, J. Taylor, Using probabilistic sampling-based sensitivity analyses for indoor air quality modelling, Build. Environ. 78 (2014) 171–182.
[23] X. Li, J. Wen, Review of building energy modeling for control and operation, Renew. Sustain. Energy Rev. 37 (2014) 517–537, https://doi.org/10.1016/j.rser.2014.05.056.
[24] S. Yigit, A machine-learning-based method for thermal design optimization of residential buildings in highly urbanized areas of Turkey, J. Build. Eng. 38 (2021), https://doi.org/10.1016/j.jobe.2021.102225.
[25] X. Li, W. Zhou, L. Duanmu, Research on air infiltration predictive models for residential building at different pressure, Build. Simul. 14 (3) (2021) 737–748, https://doi.org/10.1007/s12273-020-0685-3.
[26] W.A.Y. Mousa, W. Lang, T. Auer, W.A. Yousef, A pattern recognition approach for modeling the air change rates in naturally ventilated buildings from limited steady-state CFD simulations, Energy Build. 155 (2017) 54–65, https://doi.org/10.1016/j.enbuild.2017.09.016.
[27] C. Ding, K.P. Lam, Data-driven model for cross ventilation potential in high-density cities based on coupled CFD simulation and machine learning, Build. Environ. 165 (2019), https://doi.org/10.1016/j.buildenv.2019.106394.
[28] M.W. Ahmad, M. Mourshed, Y. Rezgui, Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption, Energy Build. 147 (2017) 77–89, https://doi.org/10.1016/j.enbuild.2017.04.038.
[29] S. Papadopoulos, E. Azar, W.-L. Woon, C.E. Kontokosta, Evaluation of tree-based ensemble learning algorithms for building energy performance estimation, J. Build. Perform. Simul. 11 (3) (May 2018) 322–332, https://doi.org/10.1080/19401493.2017.1354919.
[30] J.-S. Chou, D.-K. Bui, Modeling heating and cooling loads by artificial intelligence for energy-efficient building design, Energy Build. 82 (2014) 437–446, https://doi.org/10.1016/j.enbuild.2014.07.036.
[31] A. Rackes, A.P. Melo, R. Lamberts, Naturally comfortable and sustainable: Informed design guidance and performance labeling for passive commercial

buildings in hot climates, Appl. Energy 174 (2016) 256–274, https://doi.org/10.1016/j.apenergy.2016.04.081.
[32] Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Applying support vector machine to predict hourly cooling load in the building, Appl. Energy 86 (10) (2009) 2249–2256, https://doi.org/10.1016/j.apenergy.2008.11.035.
[33] B. Dong, C. Cao, S.E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy Build. 37 (5) (2005) 545–553, https://doi.org/10.1016/j.enbuild.2004.09.009.
[34] M. Ferreira, M. Almeida, A. Rodrigues, C. Araújo, and J. Guimarães, *PORTUGAL EPBD National report on calculation of cost-optimal levels of minimum energy performance requirements for residential buildings*. 2014.
[35] J. Feijó-Muñoz *et al.*, *Permeabilidad al aire de los edificios residenciales en España. Estudio y caracterización de sus infiltraciones*. 2019.
[36] CEN, *EN 16798-7:2017 Energy performance of buildings - Ventilation for buildings - Part 7: Calculation methods for the determination of air flow rates in buildings including infiltration (Modules M5-5)*. 2017.
[37] CEN, *Ventilation for buildings - calculation methods for the determination of air flow rates in buildings including infiltration. (EN 15242-2007)*. 2007.
[38] T. Coleman, M. A. Branch, and A. Grace, "Optimization toolbox," *Use with MATLAB. User's Guid. MATLAB 5, Version 2, Relaese II*. 1999.
[39] M. J. D. Powell, "A Fortran subroutine for solving systems of nonlinear algebraic equations," Atomic Energy Research Establishment, Harwell, England (United Kingdom). 1968.
[40] I.S. Walker, D.J. Wilson, Field Validation of Algebraic Equations for Stack and Wind Driven Air Infiltration Calculations, HVAC&R Res. 4 (2) (Apr. 1998) 119–139, https://doi.org/10.1080/10789669.1998.10391395.
[41] T.-O. Relander, S. Holøs, J.V. Thue, Airtightness estimation—A state of the art review and an en route upper limit evaluation principle to increase the chances that wood-frame houses with a vapour- and wind-barrier comply with the airtightness requirements, Energy Build. 54 (2012) 444–452, https://doi.org/10.1016/j.enbuild.2012.07.012.
[42] H.E. Feustel, COMIS—an international multizone air-flow and contaminant transport model, Energy Build. 30 (1) (1999) 3–18.
[43] J. Allegrini, J. Carmeliet, Simulations of local heat islands in Zürich with coupled CFD and building energy models, Urban Clim. 24 (2018) 340–359, https://doi.org/10.1016/j.uclim.2017.02.003.
[44] A. Hayati, M. Mattsson, M. Sandberg, Evaluation of the LBL and AIM-2 air infiltration models on large single zones: Three historical churches, Build. Environ. 81 (2014) 365–379, https://doi.org/10.1016/j.buildenv.2014.07.013.
[45] A.K. Persily, G.T. Linteris, A comparison of measured and predicted infiltration rates, ASHRAE Trans. 89 (2B) (1983) 183–197.
[46] F. Haghighat, J. Rao, P. Fazio, The influence of turbulent wind on air change rates—a modelling approach, Build. Environ. 26 (2) (1991) 95–109.
[47] M.V. Swami, S. Chandra, Correlations for pressure distribution on buildings and calculation of natural-ventilation airflow, ASHRAE Trans. 94 (3112) (1988) 243–266.
[48] H. Gough et al., Influence of neighbouring structures on building façade pressures: Comparison between full-scale, wind-tunnel, CFD and practitioner guidelines, J. Wind Eng. Ind. Aerodyn. 189 (2019) 22–33, https://doi.org/10.1016/j.jweia.2019.03.011.
[49] CEN, "EN 16798-1:2019 Energy performance of buildings - Ventilation for buildings - Part 1: Indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acous," 2019.
[50] European Collaborative Action, Guidelines for Ventilation Requirements in Buildings, Office for Official Publications of the European Community Luxembourg, EUR, 1992.
[51] T. Sağ, H. Kahramanlı Örnek, Classification rule mining based on Pareto-based Multiobjective Optimization, Appl. Soft Comput. 127 (2022), https://doi.org/10.1016/j.asoc.2022.109321.
[52] C. Du, B. Li, H. Liu, Y. Ji, R. Yao, W. Yu, Quantification of personal thermal comfort with localized airflow system based on sensitivity analysis and classification tree model, Energy Build. 194 (2019) 1–11, https://doi.org/10.1016/j.enbuild.2019.04.010.
[53] E. Mousavi, A. Bhattacharya, Event based approach for modeling indoor airflow patterns, J. Build. Eng. 51 (2022), https://doi.org/10.1016/j.jobe.2022.104244.
[54] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation," 2018.
[55] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of k-fold cross-validation, Adv. Neural Inf. Process. Syst. 16 (2003).
[56] B.G. Marcot, A.M. Hanea, What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?, Comput Stat. 36 (3) (2021) 2009–2031, https://doi.org/10.1007/s00180-020-00999-9.
[57] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Stat. Surv. 4 (2010) 40–79, https://doi.org/10.1214/09-SS054.
[58] C.M. Bishop, Pattern recognition and machine learning, Springer, 2006.
[59] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.
[60] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and regression trees, CRC Press, 1984.
[61] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer Science & Business Media, 2009.
[62] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[63] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140, https://doi.org/10.1007/BF00058655.
[64] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.
[65] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[66] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, 2018, vol. 2018-Decem, pp. 6638–6648, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85063596785&partnerID=40&md5=d6ca8cfee1067355c65e5daad4a245d5.
[67] B. Ibragimov and G. Gusev, "Minimal variance sampling in stochastic gradient boosting," in *Advances in Neural Information Processing Systems*, 2019, vol. 32, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090174574&partnerID=40&md5=aebfa421dd277d564830d4696cee7db1.
[68] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (4) (2002) 367–378.
[69] J. Fan, X. Wang, F. Zhang, X. Ma, L. Wu, Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data, J. Clean. Prod. 248 (2020), https://doi.org/10.1016/j.jclepro.2019.119264.
[70] P.W. Khan, Y.-C. Byun, S.-J. Lee, N. Park, Machine learning based hybrid system for imputation and efficient energy demand forecasting, Energies 13 (11) (2020) pp, https://doi.org/10.3390/en13112681.
[71] S. Mangalathu, H. Jang, S.-H. Hwang, J.-S. Jeon, Data-driven machine-learning-based seismic failure mode identification of reinforced concrete shear walls, Eng. Struct. 208 (2020), https://doi.org/10.1016/j.engstruct.2020.110331.
[72] J.F. Díez-Pastor, J.J. Rodríguez, C. García-Osorio, L.I. Kuncheva, Random Balance: Ensembles of variable priors classifiers for imbalanced data, Knowledge-Based Syst. 85 (2015) 96–111, https://doi.org/10.1016/j.knosys.2015.04.022.
[73] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[74] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*, 2005, pp. 878–887.
[75] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 2008, pp. 1322–1328.
[76] F. R. Torres, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "SMOTE-D a deterministic version of SMOTE," in *Mexican Conference on Pattern Recognition*, 2016, pp. 177–188.
[77] M. Mukherjee, M. Khushi, "SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features", *Applied System*, Innovation 4 (1) (2021) pp, https://doi.org/10.3390/asi4010018.