BuiltHub

DELIVERABLE D4.4: Description of a workflow on data transformation and implementation as part of the BuiltHub Road map and business case

Project acronym	BuiltHub
Full title	Deliverable 4.4: Description of a workflow on data transformation and implementation as part of the BuiltHub Road map and business case
GA no	957026
WP, Deliverable #	4, D4.4
Version	3
Date	15.09.2023
Dissemination Level	Public
Deliverable lead	RISE Research Institutes of Sweden
Author(s)	Mikael Mangold, RISE Claes Sandels, RISE Pei-Yu Wu, RISE Tim Johansson, RISE
Reviewer(s)	Marianna Papaglastra, Sympraxis Judit Kockat, BPIE Ivan Jankovic, BPIE
Keywords	Building stock analysis, applied machine learning, literature review, workflow development

This project has received funding from the EU's Horizon 2020 programme under grant agreement no 957026.

Disclaimer

The sole responsibility for the content of this publication lies with the authors. It does not necessarily reflect the opinion of the European Union. Neither the EASME nor the European Commission is responsible for any use that may be made of the information contained therein.

Table of contents

List of abbreviations

AEC	Architecture, Engineering and Construction
AI	Artificial Intelligence
ANN	Artificial Neural Network
BART	Bayesian Additive Regression Trees
BIM	Building Information Models
BSO	Building Stock Observatory
CDD	Cooling Degree Day
DW	Data Warehouse
EC	European Commission
EPC	Energy Performance Certificates
ETL	Extract Transform Load
EU	European Union
FME	Feature Manipulation Engine
GBoost	Gradient Boosting
GIS	Geographical Information System
HDD	Heat Degree Day
HMI	Human-Machine Interface
IAQ	Indoor Air Quality
IEQ	Indoor Environment Quality
kNN	k-Nearest Neighbour
Lasso regression	Least Absolute Shrinkage and Selection Operator
LCA	Life Cycle Assessment
LGBoost	Light Gradient boosting
ML	Machine Learning
MRA	Multivariate linear regression
PCA	Principal Component Analysis
PCR	Principal Component Regression
RBC	Rule-Based heuristic Control
RLC	Reinforcement Learning Control
SVM	Support Vector Machine
T4.1	Task 4.1 Specification of flexible indicators and platform information
T4.2	Task 4.2 Protocol to produce reliable figures representing EU building stock
T4.3	Task 4.3 Data organization and structure
T4.4	Task 4.4 Data transformation and clustering
T4.5	Task 4.5 Data visualization and presentation
UI	User Interface
WP4	Work Package 4 Data processing and analytics
XGBoost	Extreme gradient boosting decision-tree

Executive summary

The execution of *Task 4.4 Data transformation and clustering* is part of the Horizon 2020 (H2020) project BuiltHub, a European project aiming to create a dynamic EU building stock knowledge hub. By linking the potential data sources and the building-related policy and business, the project explores the benefits of developing community-enhanced data-driven applications (1). The outcome of the project is expected to contribute to European energy efficiency policies and key directives, such as Energy Performance in Building Directive (EU) 2018/844, Energy Efficiency Directive (EU) 2018/2002, Renewable Energy Directive (EU) 2018/2001, and Renovation Wave (COM) 2020/662.

To support the data-driven policy formulation for the building sector, Working Package 4 focuses on conducting the data processing and analytics tasks for the BuiltHub, including indicators specification, protocol creation, data structure systemization, as well as data transformation and visualization. The ultimate objective of Task 4.4 was to transform and cluster available datasets from Builthub. The periodic goals are two-folded: (1) screening the examples in literature to guide potential analyses in BuiltHub, particularly in the later work for Task 4.4 and 4.5, (2) providing the evaluators of the BuiltHub project with a robust background work that addss the feasibility when selecting the work to focus in the task. To serve the purposes mentioned above, Deliverable 4.4 builds upon the results of *Task 4.1 Specification of flexible indicators and platform information* and further investigates relevant machine learning methods for future building stock analysis. Furthermore, Deliverable 4.4 exemplifies what methods and tools can be used by partners to use machine learning for developing features using machine learning that can be used as estimates for assessing energy efficiency strategies as part of the BuiltHub roadmap.

A mixture of theoretical background regarding building stock analysis and practical research applications are introduced to orient the BuiltHub work in the current domain development. After scientific positioning, a standard workflow is explicitly proposed to navigate the BuiltHub building stock analysis, which can be used to assess the requisites in each step. Finally, the technical suggestions to overcome the predicaments of the BuiltHub datasets are described for future work.

The deliverable is structured as follows: Section 2 discusses the research field of building stock analysis in general, including the evolution of the subject and previous approaches for conducting analysis. Thereafter, Section 3 defines the scope of Deliverable 4.4, explaining the relationship between theory, research examples, and an applied machine learning workflow. Section 4 concerns the theoretical part of machine learning application in building stock analysis. A brief introduction to machine learning techniques is given by exhibiting the relationship between data science, artificial intelligence, and machine learning. The strengths and the weaknesses of the machine learning models are presented to facilitate the model selection. Built upon the theoretical background, a comprehensive review of the building stock research using machine learning techniques is performed. Through demonstrating practical examples, the aim is to show the possibilities of building stock analysis and further research opportunities for pre-defined thematic areas. Critical information regarding input datasets, the aggregation level, and the building types is also identified in Section 4.

Section 5 illustrates what steps partners can take to incorporate machine learning in their analyses as part of the BuiltHub roadmap. It includes the workflow development for the applied machine learning loop with five sequential steps to work with data-driven problem-solving for partners, data merging techniques, and some analyses. In section 5 three examples are included to further illustrate what steps partners should take. The examples are describing different levels of applying ML in the prediction of hazardous materials in the existing building stock. Despite different data input and research purposes, both examples and the Swedish Pilot case in section 6 show how specific information can be added to the national building database using machine learning methods.

Section 6 describes the Swedish Pilot where the Swedish authorities were assisted in developing a building-specific national energy efficiency strategy. Machine learning was used to predict building types for which costs and efficacies for energy efficiency renovation measures had been developed. Finally, a last section has been added to link the Swedish case with exploitation possibilities as the BuiltHub project is coming towards an end. There are several opportunities for exploitation of BuiltHub results that could explored.

Introduction

Research on building stock is by its nature interdisciplinary and complex. Since the built environment is heterogeneous and regional-dependent, the building stock analysis concerns a variety of economic, social, and environmental aspects at different aggregation levels. The building stock analysis can be approached from several perspectives, and the thematic development is uneven. Driven by the need to understand energy consumption and the increasing necessity to refurbish post-war buildings, the pursuit of economic efficiency opened the prelude of building stock modelling (2). Besides, the urge of climate change accelerates the focus of environmental sustainability of building stock management (3). Compared to the economic and environmental-oriented building stock research, the social dimensions regarding affordability, equality, and segregation among the occupants are underdeveloped (4). Given the obscure nature and comprehensiveness of the subject, concrete examples and defining indicators are used to measure the sustainable development of building stock study (5).

Simultaneously, advanced development of methods and tools offer new opportunities to study the building stock quantitatively and qualitatively. For instance, remote sensing and geographical information system (GIS) enable spatial investigation with multiple data input formats. Other data analytic techniques and programming languages, such as Tableau, R, Matlab, Feature Manipulation Engine (FME) and Python, also offer new visualization possibilities for urban analytics. Implementing the digital tools may better capture the dynamics of the building stock and provide more reliable models to describe its composition using statistical results. However, significant challenges remain in collecting digital data for specific research purposes.

The diverse research spectrum of the subject and insufficient fundamental data results in slow advance (2). To overcome the challenges of partial analysis, the emergent building stock research tried to connect different types of information to conduct a more comprehensive and in-depth study. Machine learning comes into play to leverage the limited known examples to predict the unknown instance at a large scale. Other advantages of applying machine learning methods include rapid prediction iteration, cost-efficient hypothesis testing, change monitoring for time series data, etc. Consequently, it can be used to effectively validate field data such as energy performance certificates (EPC) and update the change of registered data due to the shift of renovation strategy policies (6).

Scope

This deliverable aims to describe what machine learning can be within building stock analysis. Figure 1 below illustrates the structure of Deliverable 4.4 and the association between each element. By reviewing the applied machine learning approaches in previous building stock research, a comprehensive view of data analytics' potential in various thematic topics in their respective theoretical background has been created. Afterwards, an in-depth discussion about available examples, along with the applied machine learning loop workflow, was described. Lastly, the possible analysis and challenges to work with the BuiltHub building stock data were presented.

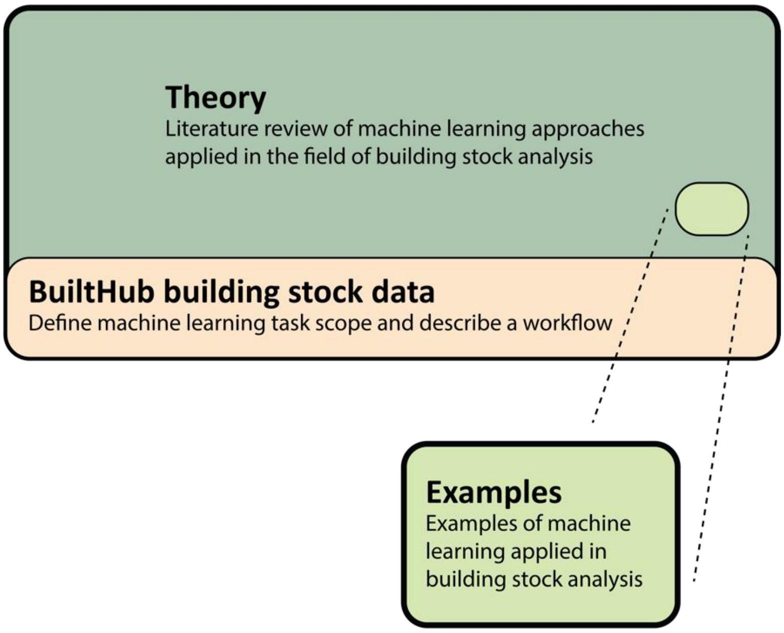

Theory
Literature review of machine learning approaches applied in the field of building stock analysis
BuiltHub building stock data
Define machine learning task scope and describe a workflow
Examples
Examples of machine learning applied in building stock analysis

Figure 1. Scope of Deliverable 4.4.

To guide building stock analysis, the thematic areas of the BuiltHub data structure, along with the indicators and the datasets available in the project were created in Table 1. The referenced indicators are adopted from Task 4.1, produced according to the thematic areas to present the BuiltHub platform information. Topics 1 to 5 concern existing Building Stock Observatory (BSO) areas with their respective thematic areas, including *energy*, *building stock*, *building characteristics*, *certification*, and *finance*. Topics 6 to 9 are new areas included in BuiltHub to comprehend the building stock analysis, such as *indoor environment quality (IEQ)*, *climate*, and *smart-grid ready buildings*. According to the data indicators from T4.1 and the matrix from T4.2, a majority of the BuiltHub datasets concerned *1.1 Energy consumption* and *2.1 Building stock characteristics*. These datasets mainly originated from EUROSTAT with similar data structures and categorical units. However, there is a lack of data in the thematic areas *3.2 Technical building systems*, *6.2 Indoor air quality, 6.3 Natural lighting, 8.1 Loading stations*, and *9.1 Smart-grid ready buildings*. As highly aggregated data from the EU member countries were compiled from 30 different sources, common indicators were developed for assessment purposes in comparative studies. The indicators for *1.1 Energy consumption* concern, for instance, total or specific energy consumption for the residential sector in total, per building or dwelling, or per square meter. Similar indicators were created for *1.3 Energy market, 3.2 Technical building systems, 6.1 Comfort, 6.2 Indoor air quality, 6.3 Natural lighting*, and *9.1 Smart-grid ready buildings* to describe statistical features of the data such as count, mean values, and share.

Table 1 Thematic structure of the BuiltHub building stock analysis with indicators and available datasets.

Topic	Thematic areas	Indicators	Available dataset *
Existing Building Stock Observatory (BSO) areas			
1. Energy	1.1 Energy consumption	1. All-end-uses Total Energy consumption for the residential sector. 2. Space heating Total Energy consumption for the residential sector. 3. Domestic hot water Total Energy consumption for the residential sector. 4. Electricity consumption of lighting for the residential sector. 5. Space cooling energy consumption for the residential sector. 6. Space heating Energy consumption of single-family residential sector. 7. Total energy consumption per building. 8. Total Energy consumption per dwelling in the residential sector. 9. Space heating energy consumption per dwelling for the residential sector. 10. Water heating energy consumption per dwelling for the residential sector. 11. Lighting energy consumption per dwelling for the residential sector. 12. Energy consumption per m² for the residential sector. 13. Space heating energy consumption per m² for the residential sector. 14. Space cooling energy consumption per m² for the residential sector.	1;6;8;12;14;15;16;22
	1.2 Energy poverty	NA	17;23;25;26
	1.3 Energy market	1. Average energy price of natural gas. 2. Average energy price of fuel oil. 3. Average energy price of coal. 4. Average energy price of electricity. 5. Average energy price of biomass.	24
2. Building stock	2.1 Building stock characteristics	1. The total number of dwellings.	1;2;3;5;6;7;10;11;17;19;22
	2.2 Building renovation	NA	13

3. Building characteristics	3.1 Building shell performance	NA	21
	3.2 Technical building systems (incl. smart meters)	1. Share of dwellings with condensing boilers 2. Share of dwellings with solar heating system 3. Share of residential dwellings with a combi boiler 4. Number of dwellings with heat pumps 5. Number of dwellings with the heating on electricity 6. Number of residential dwellings with electric heaters (not heat-pump) for water heating	NA
	3.3 Nearly zero-energy buildings	NA	4;5;13
4. Certification	4.1 Certification	NA	4;18
5. Finance	5.1 Financing	NA	20
New areas suggested by BuiltHub			
6. Indoor Environment Quality (IEQ)	6.1 Comfort	1. mean HDD (per country/region). 2. mean CDD (per country / region) 3. share of buildings that can keep the house warm 4. share of households with a leaking roof 5. share of households with leaking walls 6. share of households with leaking windows 7. share of offices with individual temp. control	27
	6.2 Indoor air quality	1. share of office space with ventilation	NA
	6.3 Natural lighting	1. share of office space with natural lighting	NA
7. Climate	7.1 CO2 emission	NA	28
	7.2 Climatic conditions**	NA	29
	7.3 Solar radiation	NA	30
8. E-mobility	8.1 Loading stations	NA	NA
9. Smart-grid ready buildings	9.1 Smart-grid ready buildings	1. total numbers of installed meters 2. share of installed meters in the existing buildings	NA

* The 30 datasets listed below are currently accessible in the BuiltHub project. The column Available Dataset corresponds relevant datasets and thematic areas.

Dataset 1: Horizon 2020 HotMaps project: Building stock analysis
Dataset 2: IEE TABULA project: Typology Approach for Building Stock Energy Assessment
Dataset 3: IEE EPISCOPE project: Focus of building stock monitoring
Dataset 4: IEE ZEBRA2020 project: Nearly Zero-Energy Building Strategy 2020

Dataset 5: IEE ENTRANZE project: Policies to Enforce the TRAnsition to Nearly Zero Energy buildings in the EU27

Dataset 6: H2020 ODYSSEE - MURE project: Comprehensive monitoring of efficiency trends and policy evaluation in EU countries, Norway, Serbia and Switzerland.

Dataset 7: FP7 CommONEnergy Project: building stock

Dataset 8: JRC IDEES 2015

Dataset 9: SET-Nav - Strategic Energy Roadmap

Dataset 10: H2020 ExcEED Project: building stock data

Dataset 11: FP7 iNSPiRe project: building stock analysis

Dataset 12: Energy consumption and energy efficiency trends in the EU-27+UK for the period 2000-2016 - FINAL REPORT

Dataset 13: Comprehensive study of building energy renovation activities and the uptake of nearly zero-energy buildings in the EU - FINAL REPORT

Dataset 14: EUROSTAT: Final energy consumption in households

Dataset 15: EUROSTAT: Final energy consumption in households by fuel

Dataset 16: EUROSTAT: Disaggregated final energy consumption in households

Dataset 17: ZENSUS 2011

Dataset 18: DPE - Diagnostic de Performance Energetique

Dataset 19: Towards a sustainable Northern European housing stock - Sustainable Urban Areas 22

Dataset 20: DEEP - De-risking Energy Efficiency Platform

Dataset 21: Energy consumption and efficiency technology measures in European non-residential buildings

Dataset 22: Dataset of the publication: Europe's Building Stock and Its Energy Demand: A Comparison Between Austria and Italy

Dataset 23: National Housing Census: European statistical System

Dataset 24: Energy prices in 2019 - Household energy prices in the EU

Dataset 25: EUROSTAT: GDP per capita in PPS

Dataset 26: EUROSTAT: Population on 1 January by age, sex and NUTS 2 region

Dataset 27: EUROSTAT - Cooling and heating degree days

Dataset 28: EDGAR (Emissions Database for Global Atmospheric Research) CO2 Emissions

Dataset 29: CORDEX - Regional climate model data on single levels for Europe

- Dataset 30: PVGIS - Photovoltaic Geographical Information System

** Climatic conditions refer to the climate parameters relevant for buildings' energy demand, such as temperature, heat degree day (HDD), cooling degree day (CDD), and wind speeds.

Applicability of machine learning approaches

After delineating the thematic areas and indicators in Section 3, a literature review concerning the state-of-art of applied machine learning research in the building stock studies helps position the subsequent analysis in theoretical contexts. Section 4 investigates applicable machine learning methods for different research purposes and the requirements for conducting analysis. Firstly, a general introduction of machine learning techniques related to other decision systems is given using a hierarchical diagram, see Section 4.1. Secondly, by reviewing building stock research using machine learning techniques, relevant techniques, aggregation levels, and datasets are organized in a matrix in Section 4.2. The last part, Section 4.3, discusses the criteria for machine learning model selection and presents their strengths and weaknesses.

Introduction of machine learning techniques

To understand how machine learning relates to other computational domains or decision-making systems, Figure 2 is used to clarify the hierarchical relationship among various terms. On top of the diagram, data science encompasses the sub-domains of "scientific calculation," "artificial intelligence," and "human-machine interface." Data science refers to the science of integrating data by applying mathematics, statistics, computer science, and domain knowledge. Using scientific methods, processes, algorithms, machinery systems extract knowledge and insights from structured and unstructured data (7,8). A standard data science process involves data collection, processing, exploration, model building, and results communication (9).

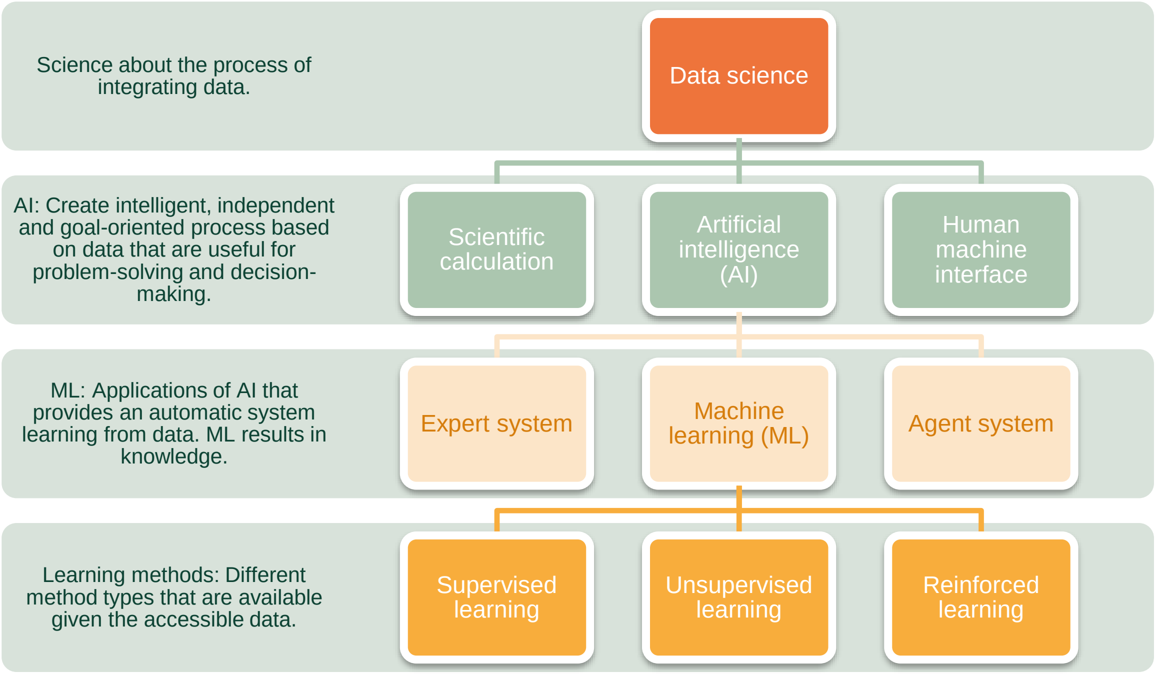

Science about the process of integrating data.
Data science
AI: Create intelligent, independent and goal-oriented process based on data that are useful for problem-solving and decision-making.
Scientific calculation
Artificial intelligence (AI)
Human machine interface
ML: Applications of AI that provides an automatic system learning from data. ML results in knowledge.
Expert system
Machine learning (ML)
Agent system
Learning methods: Different method types that are available given the accessible data.
Supervised learning
Unsupervised learning
Reinforced learning

Figure 2. Relationship between data science, artificial intelligence, and machine learning.

In traditional building science, scientific calculation plays an essential role in building physics, structural engineering, building materials, etc., to predict and enhance buildings' performance and sustainability (10). Computational tools are frequently employed in the design phase to simulate building performance based on the input information about the built-in systems. The output, such as energy use, ventilation efficiency, heat radiation, etc., is evaluated for design optimization (11). Since the computational tools and simulation models are developed, their accuracy depends on the expert's knowledge, assumption setting, and program validation (12). Hence, the limitations of the simulation approach present the requirements of comprehensive and concrete environmental assumptions, which is usually possible in a small-scale empirical study.

Another way of harnessing machinery computation power is through a human-machine interface (HMI) using input and output hardware. Originated from the user interface (UI) in industrial design, the ideal UI targets optimizing operation or control from humans to machines while leveraging the feedback from machines to assist the decision-making process. Operational systems and programming languages are examples of HMI; such functions are required to be pre-programmed explicitly by humans (13). In the last decades, artificial intelligence (AI) was developed to overcome the limitations of the high involvement of human factors in programming.

AI describes the process of creating intelligent, independent, and goal-oriented learning based on the data that are useful for problem-solving and decision-making. In other words, AI can be regarded as intelligent agents who perceive their environment and take actions to maximize the chance of succeeding in the defined tasks (14). The conventional tasks or goals for AI contain knowledge representation and reasoning, automatic planning and scheduling, machine learning, natural language processing, and machine perception (15). In the last decades, a tendency of the uptake of AI in the Architecture, Engineering and Construction (AEC) industry was observed substantially. The trend can be reflected by a growing number of research works in the AI-related field, which is due to the complex and difficult problems faced by the industry, such as building lifecycle evaluation, performance assessment, robotic automation application, and so on (16). The adoption of AI introduces opportunities to the long-standing challenges by optimizing the scientific calculation and simulation and streamlining the traditional HMI process. The maturity of algorithm development and computational power, along with recognizing the benefits of AI and high data availability in the recent decade, led to an upsurge in applied studies in the subject, including building life cycle (17), construction and demolition waste (18), indoor air quality research (19) and so on. Despite being preliminary and heuristic, the developed applications offer new perspectives and advance the conventional practice.

As an AI application, machine learning provides an automatic system learning from data and generates knowledge (20). Machine learning models are built based on the known instances in training data to predict unknown properties without explicitly programming (21). On the other hand, data mining employs similar methods like machine learning to discover unknown properties in the data. Due to the high flexibility, machine learning is widely adopted in several fields for pattern recognition, including medicine, business, and building sciences, when developing conventional algorithms is not a viable option (22). In comparison, expert systems and agent systems are as common as machine learning in building sciences despite computer aids. Expert systems utilize if-then rules to simulate decision-making capability rather than taking in procedural codes (23). An expert system consists of the inference engine and the knowledge base. The inference engine applies the rules to the known facts to deduct new facts, where rules and facts are saved in the knowledge base (24). On the other hand, agent systems rely on the intelligent agent to directly interact with the physical or virtual environment. Multiple agents are equipped with a sensor to receive signals, process signals with an intelligent program, and react to the environment with specific goals (25). In short, the abovementioned machine learning, the expert system, and the agent system belong to artificial intelligence.

Different methods have been developed in the domain, given accessible data and prediction purposes according to the learning requirements. Three major types of machine learning can be distinguished, i.e., supervised learning, unsupervised learning, and reinforcement learning. Supervised learning exploits the known features on data to generate insights for the unknown examples; two primary categories of supervised learning problems are “regression” and “classification.” Conversely, no data labels are given in unsupervised learning; thus, this type of learning deals with clustering and transformation for density estimation. Lastly, reinforcement learning is applied in a dynamic situation using intelligent agents to attain pre-defined goals through maximizing rewards. The application areas for reinforcement learning can be seen in engineering subjects, where optimized resource management and control are desired. Supervised and unsupervised learning are prevalent in building sciences applications, whereas reinforcement learning is usually studied in operations research, control theory, multi-agent systems, etc. To understand the match between machine learning models, datasets, and the purpose of analysis, Section 4.2 introduces the criteria for algorithm selection.

Criteria for machine learning model selection

The criteria for selecting machine learning models depend highly on the data types and the hypothesis formulation. Overall, various machine learning models present different strengths and weaknesses for analysis. Table 2 lists the common model types, descriptions, and characteristics for four learning problems: supervised learning, unsupervised learning, reinforcement learning, and deep learning. First, supervised learning can be roughly distinguished to regression models, classification models, and decision tree models. Regression models assume a linear relationship between dependent and predictive variables and handle mostly continuous values. Exceptions are generalized linear regression, which can process both continuous and categorical variables and transform non-linear problems. Also, regularized linear model, partial least squares, and principal component regression are suitable for high dimensional datasets, i.e., the number of observations is fewer than the numbers of variables, as well as to handle multicollinearity problems, i.e., the correlation between predictive variables that causes sensitive and less precise coefficient estimates.

On the other hand, classification models employ linear and non-linear classifiers to deal with discrete values for binary, multi, or nominal dependent variables. They are used to predict probability with robust and incremental learning capability. Due to high flexibility and easy interpretation, support vector machine and k-NN is frequently used in building stock analysis for pattern identification, such as building type and design feature prediction. The former separates the data groups by projecting data in space and determining their categories based on the gap, while the latter leverages incremental learning to estimate the likelihood of data groups. Lastly, the decision tree family is a decision support model for continuous and categorical variables. No variable selection is required, and the capability to deal with missing data is the most significant advantage. However, the decision tree models tend to be overfitting, and modeling error occurs for learning noises in the training dataset and compensates the generalizability for new datasets, thus needing careful tunning.

Table 2: Strengths and weaknesses of the applied machine learning models adopted from Wei et al. (19).

Category	Model	Description	Strength	Weakness
Supervised learning				
Regression models	Multiple linear regression	Linear regression models for *continuous*, dependent variables	1. Determine predictive variables 2. Detect outliers 3. Flexible variables selection methods, i.e., stepwise, forward, backward algorithms	1. Sensitive to outliers 2. Requires more observations than variables 3. The risk of multicollinearity
	Generalized linear regression	Linear regression models for *continuous and categorical* variables. Response variables with errors and not normally distributed are allowed	1. Transformation of non-linear problems to linear problems 2. Flexible variables selection methods, i.e., stepwise, forward, backward algorithms	1. Sensitive to outliers 2. Requires more observations than variables 3. The risk of multicollinearity
	Regularized linear model (LASSO regression, Ridge regression)	Linear regression models for *continuous* variables that regulate or shrink the coefficient estimates toward zero	1. The number of observations can be lower than the number of variables 2. Prevent overfitting 3. Perform variable selection 4. Multicollinearity is handled	1. LASSO regression: restricted number of selected variables 2. Ridge regression: Unable to classify the important level of variables
	Partial least squares	Linear regression models for *continuous and categorical* variables that project variables to a new space	1. The number of observations can be lower than the number of variables 2. Can have more than one dependent variable	1. The number of components is chosen based on cross-validation and cross-validated R^2 (Q indicator)

			3. Deals with missing data 4. Multicollinearity is handled	
	Principal component regression (PCA)	Linear regression models for *continuous* variables based on principal component analysis	1. The number of observations can be lower than the number of variables 2. Multicollinearity is handled	1. Dependent variable is out of consideration when choosing the principal components
Classification models	Logistic regression	Linear classifiers for the discrete variable as the output is transformed to log-odds	1. Estimate the probability 2. Make no assumptions about distributions of classes in feature space 3. Provide coefficient size and direction of the association	1. Require average or no multicollinearity between predictive variables 2. Can be overfitted in high dimensional datasets, i.e., number of observations is less than the number of variables
	Support vector machine (SVM)	Linear and non-linear classifiers that projected the categorical data in space and determined their categories based on the gap	1. Capable to handle high dimensional data	1. Hard to choose the penalty variable 2. Hard to choose the kernel
	Naïve Bayes	Linear and non-linear classifiers that use Bayes' theorem to calculate the probability	1. Capability to handle high dimensional data 2. Fast processing 3. Robust and incremental learning	1. No variable selection 2. No explicit model 3. Strong naïve independence assumption between the features
	k-NN	Linear and non-linear classifiers that estimate the likelihood of data group based on close proximity	1. Simple and incremental learning	1. Slow processing 2. Hard to tune the model and interpret results 3. Variables need to be at similar scales
Decision tree models	Decision tree	Linear decision support models that take in	1. No need for prior variable selection	1. Requires a large amount of data to train

		continuous and categorical variables and present possible consequences	2. Deals with missing data	2. Have overfit tendency
	Gradient boosting tree	Linear and non-linear decision support models for continuous and categorical variables that combine decision tree algorithms and boosting methods	1. Accuracy is improved 2. Mediate overfitting	1. Requires model-tuning 2. Slow processing due to cannot be parallel processing
	Random forest	Linear and non-linear decision support models for continuous and categorical variables that combine the prediction results of multiple trees	1. No need for prior variable selection 2. Accuracy is improved 3. Mediate overfitting 4. Parallel processing	1. Hard to interpret the predictors 2. Multiple parameters to tune, i.e., number of features, trees, and minimum sample leaf size
Unsupervised learning				
Clustering	Hierarchical clustering, distribution-based clustering, density-based clustering, k-means clustering	A grouping method for similar data characteristics	1. Optimize between intraclass and interclass variance	1. Determine the number of groups 2. Variables needs to have similar scales 3. Results may vary on the algorithm and tunning method
Transformation	Principal component analysis (PCA), multiple correspondence analysis (MCA)	An information projection method to study the similarities between individuals and variables in a multidimensional space	1. Deal with a large amount of data	1. Variables needs to have similar scales
Reinforcement learning				

Control learning	Monte Carlo, Tabular Q-learning, Batch Q-learning	Optimal control for the closed-loop problems	1. Consider the whole problem and prevent local optimization	1. Trade-off between exploration and exploitation
Deep learning				
Artificial neural networks	Feed-forward back-propagation network, cascade correlation	(Supervised learning) An interconnected linear and non-linear neuron-like structure that classifies continuous and categorical variables	1. Deal with missing data 2. Parallel processing 3. Apply to various types of problems, i.e., image and sound recognition, text, time series	1. Require strong computation power 2. Difficult to tune the model 3. Hard to understand the behavior of the network
	Autoencoder neural network, self-organizing map	(Unsupervised learning)		

The objective of unsupervised learning is to capture the information structure of the datasets without available labels to verify the prediction results. They are used to discover clusters and detect outliers, extract, and compress and summarize the data. Two major types of unsupervised learning exist, i.e., clustering and transformation. Clustering divides data points into different groups based on their similarity. The number of clusters and the linkage criteria should be defined to initiate the iterative, bottom-up approach. Another type of unsupervised learning is transformation, which describes the process for extracting or computing information. Factorial analysis, such as Principal component analysis (for numerical variables) and multiple correspondence analysis (for categorical variables), are standard techniques for dimensionality reduction. The two-dimensional approximation is created to capture most of the variation in the dataset. Normalization is needed before clustering and transformation to scale the variables evenly. Both methods can visualize data points relationships in a dataset and can be used for compression to identify features for supervised learning.

Reinforcement learning describes goal-oriented agents that are interacting with an uncertain environment constantly (26). The agent's action affects the options and uses its experience to improve opportunities later. Thus, it seeks to balance the trade-off between exploration and exploitation. In short, reinforcement learning is studied in optimal control theory and is widely adopted in engineering subjects, including building system control. Deep learning is a subfield of machine learning that simulates neural networks with representation learning. Representation learning, also called feature learning, automatically detects features or classification from raw data (27). Deep learning encompasses supervised and unsupervised learning and progressively uses multiple layers to extract high-level features from data (28). Deep learning applications have been seen in automatic speech recognition, image recognition, natural language processing, recommendation system, etc. In the next section, the status quo of applied machine learning in the building stock research will be reviewed with an explicit focus on supervised, unsupervised, reinforcement learning, and deep learning.

Review of the building stock research using machine learning techniques

Nine major topics are identified in the building stock analysis and expanded to several thematic areas, structured in Table 3. The first five topics are derived from the existing BSO areas: *1. Energy*, *2. Building stock*, *3. Building characteristics*, *4. Certification*, and *5. Finance*. The new topics suggested by BuiltHub are *6. Indoor environment quality (IEQ)*, *7. Climate*, *8. E-mobility*, and *9. Smart-grid-ready building*. The matrix below maps the previous building stock research using machine learning techniques in different thematic areas. By demonstrating their research purpose, machine learning method, aggregation level, and input data can help delineate the status quo of the building stock study. These research examples are used to orient the analysis of the available BuiltHub datasets.

Table 3: Summary of the building stock studies using machine learning methods categorized according to thematic areas. Each machine learning method and its abbreviation are explained shortly below the table.

Thematic areas	Ref.	Research purpose	Method*	Aggregation Level/ Building category	Prediction dataset	Training dataset
1.1 Energy consumption	(6)	Predicting heat demand indicators from the magnitude of registered buildings for EPC validation	ANN	Regional/ Residential buildings	The same dataset as training data	Energy performance certificates
1.2 Energy poverty	(29)	Categorizing energy poverty risk based on income and energy expenditure	XGBoost	National/ General buildings	The same dataset as training data	Data of house value, ownership, age, household size, average population density, household income
1.3 Energy market	(30)	Developing adaptive updating forecast system for dynamic energy market	ANN	Regional buildings	The same dataset as training data	Numerical weather prediction data, Historical power data
2.1 Building stock characteristics	(31)	Predicting the presence of hazardous materials	Statistics	Regional/General buildings	National building registers,	Environmental inventories

					Energy performance certificates	
2.2 Building renovation	(32)	Predicting building features for energy efficiency strategies	SVM, logistic regression	National/ Multifamily houses	National building registers, Energy performance certificates	Building observations from Google Street Views
3.1 Building shell performance	(33)	Benchmarking energy performance of existing residential buildings' envelopes	PCA, PCR, MRA, Fuzzy C-Means clustering	Regional/ Residential buildings	Monthly recorded climate and energy consumption data from the household community	Field survey from infrared thermography
3.2 Technical building systems (incl. smart meters)	(34)	Leveling the features importance for classifying both the building use type and performance	Supervised classification with time series analysis	Not specific/ Non-residential buildings	Building performance classification and characterization	Building electrical meter data
3.3 Nearly zero-energy buildings	(35)	Comparing modeling approach for low energy building systems	RLC, ANN	Individual building/ Residential LowEx buildings	Model comparison with rule-based heuristic control	Weather data, global solar irradiation, ambient air temperature
4.1 Certification	(36)	Proposing a new method for building energy performance benchmarking	K-means clustering	National/ General buildings	National benchmarks for validation	Multidimensional building features
5.1 Financing	(37)	Proposing an automated property valuation model	GBoost, Genetic algorithm optimization	National/ General buildings	The same dataset as training data	Traded houses property data from Github public machine learning datasets, data extraction

						from BIM models
6.1 Comfort	(38)	Achieving optimal control of HVAC and window systems for natural ventilation	Reinforcement learning	Citywide/ Residential buildings	Model comparison with rule-based heuristic control	Hourly weather data; Building parameters
6.2 Indoor air quality	(39)	Predictive mapping of indoor radon concentrations	Random forest, k-medoids clustering, BART	National/ General buildings	Lithological unit	Indoor radon concentration measurement
6.3 Natural lighting	(40)	Recommending an occupant-customized luminous environment	KNN, decision tree, random forest, SVM	National/ General buildings	The same dataset as training data	Luminous environment information from lifelog data
7.1 CO_2 emission	(41)	Predicting costs and CO_2 emission in the integrated energy-water optimization model	SVM, ANN, linear/ lasso/ ridge/ elastic net/ bagging regression, GBoost, XGBoost, light GBoost, extra trees, random forest	National/ General buildings	The same dataset as training data	Data of hybrid RWH-GR systems, ground source heat pumps, energy consumption in smart – nZEB buildings, energy producing systems, energy storage system, energy export and import to network, costs data, weather data
7.2 Climatic conditions	(42)	Predicting multiple building energy loads and BIPV power production	ANN, SVM, neural network	General buildings	Simulated energy data	Weather data, simulated building operating,

						and energy data
7.3 Solar radiation	(43)	Forecasting solar irradiation and load power consumption	ANN	District/ General buildings	Solar irradiation	Load power consumption (Load, PV system, EVs, and power grids)
8.1 Loading stations	(44)	Simulating and optimizing a PV-based energy system integrated with onsite battery and electric vehicles	GBoost	Individual/An existing residential buildings	PV production and curtailment	Sensors and measurements installed in the houses and PV/T panels, weather data
9.1 Smart-grid ready buildings	(45)	Assessing the performance of control algorithms for implementing demand response strategies	Rule-based control, Smart algorithm	Not specific/ Residential buildings	Electricity price predictor	Real-time smart meter data, weather data

* Abbreviations and short definitions for various machine learning methods:

- ANN (Artificial Neural Network): A information processing algorithm that simulates human neurons and forms the networks of the input layer, hidden layer, and output layer.
- XGBoost (Extreme gradient boosting decision-tree): A decision tree-based ensembled algorithm applies a gradient boost network to improve the prediction performance for tabular data.
- SVM (Support vector machine): A classifier that projects the data points in space and determines their categories based on the gap.
- Logistic regression: A linear classifier that transforms output to log-odds and generates predictive probability.
- PCA (Principal component analysis): An information projection method to study the similarities between individuals and numerical variables in a multidimensional space.
- PCR (Principal component regression): A regression-based PCA to estimate the unknown regression coefficients in a linear regression model.
- MRA (Multivariate linear regression): A linear regression model studies the correlation of a dependent variable and multiple independent variables to predict the output.
- Fuzzy C-Means clustering: a clustering technique that builds fuzzy partition from data by assigning each data point membership in each cluster center.
- RLC (Reinforcement learning control): A reinforcement learning for solving optimal control problems.
- k-means clustering: A clustering technique that partitions data into k clusters in which each observation belongs to the group with the nearest mean and minimizes the total squared error.
- GBoost (Gradient boosting): An iterative technique to combine weak learners into a single strong learner to optimize prediction performance.
- Genetic algorithm optimization: A random-based evolutionary algorithm to find the optimal solutions by introducing a gradual change.

- Random forest: An ensemble learning method based on combining the outcomes from multiple decision trees.
- k-medoids clustering: A clustering technique that minimizes the sum of dissimilarities between points in the same cluster and designates a point as the center of each group.
- BART (Bayesian additive regression trees): A similar ensembled technique to GBT; however, a prior in the Bayesian approach is used to sum up the contribution of each decision tree.
- kNN (k-Nearest Neighbour): A classifier that estimates the likelihood of data group based on close proximity.
- Decision tree: A linear decision support model that uses rules to separate features and generate predictive outcomes.
- Lasso regression (Least absolute shrinkage and selection operator): A regularized linear regression that applies shrinkage to reduce overfitting and add feature selection.
- Ridge regression: A regularized linear regression that shrinks coefficients to prevent multicollinearity.
- Elastic net regression: A regularized linear regression that linearly combines the L1 and L2 penalties of lasso and ridge regression.
- LGBoost (Light gradient boosting): An boosting ensembled framework that advances XGBosst at the leaf level to improve the computational speed and accuracy.
- Bagging regression: An ensembled estimator that trains the regressors on random subsets of the data and aggregates their respective prediction results.
- RBC (Rule-based heuristic control): A system that operates according to the human-made rules to handle data.

Energy consumption is the most studied area within building stock analysis with highly accessible data and the increasing need for renovation strategy. Various machine learning techniques, including regression and deep learning, were explored for data validation and prediction of energy performance certificates (EPCs). Khayatian et al. (6) estimated the heat demand indicator by testing the optimal feature combination and model selection. Their finding showed that processing only twelve variables from EPCs with the artificial neural network can achieve high prediction accuracy. In comparison to energy consumption, literature for energy poverty and the energy market are limited. These socio-economic parameters are broad and hard to quantify or monitor, making them challenging to address with policy measures (29). Longa et al. (29) employed extreme gradient boosting to predict energy poverty risk with socio-economic data as input. The results showed that machine learning could support evidence-based policymaking for the complex mechanism of energy affordability. Applied machine learning for the energy market has a different focus on dynamic forecasting. The deep neural network was explored to achieve an adaptive, updating forecasting system with the sequential power data stream (30). The objective is to create an intelligent local energy market from a systemic perspective, yet no direct connection to the building stock analysis has been observed.

Analysis regarding building stock involves building stock characteristics and building renovation. Usually, national building registers are used as a foundation and adding specific field data to complete certain research tasks. For instance, environmental inventories were compiled and merged with the building registers and EPCs data to predict the presence of hazardous building materials by Wu et al. (31). The feasibility of determining specific energy retrofit strategies was investigated by adding building characteristics to the EPCs database by Platten (32). Through observing buildings from Google Street Views, insufficient details on building type and suitability for additional façade insulation can be complemented. The prediction results from the supervised machine learning were accurate enough to improve estimations of national energy-saving potential.

Furthermore, the thematic areas related to building characteristics are three-fold, defined by T4.2: building shell performance, technical building systems, and nearly zero-energy buildings. Analyses related to the topic emphasize enhancing the technical performance of construction parts or installed systems. Wang et al. (33) proposed a new approach of using multivariate linear regression with principal component analysis to benchmark the energy performance of building envelopes. Their validation results with infrared thermography indicated that the method is preferable to the traditional statistic rating method using average energy consumption of buildings to handle the multicollinearity risk with the high dimensional dataset. Exploiting the smart meter data from non-residential buildings, Miller (34) analyzed essential temporal features for classifying various building performance attributes concerning the primary use of a building and the level of building performance. By adopting time-series analysis for load clustering, the interpretability of performance classification and customer segmentation were explored. To reduce engineering cost and the return-of-investment period for setting up the systems of low exergy buildings, Yang et al. (35) compared the performance of reinforcement learning control (RLC) approach against the conventional rule-based control (RBC) method. In general, RLC outperformed RBC concerning PV/T and the complete building control with respect to optimal heat supply, temperature adjustment, and ground heat compensation.

Certification and financing of building stock are less studied, and merely a few applied machine learning papers are available. Clustering was used to overcome the limitations of the benchmarking program by considering the impacts of the multiple features of energy performance in buildings. Gao and Malkawi (36) showed that multidimensional clustering could facilitate energy evaluation among different types of buildings. Conversely, the real estate sector adopts the AI-enhanced automated valuation models to improve its low transparency, inaccuracy, and inefficiency in property valuation rather quickly. Su et al. (37) presented an integration framework of building information models (BIM) and machine learning for automated property valuation. The aids of machine learning enable comprehensive data interpretation, and at the same time, improve information exchange between AEC projects and real estate activities.

Building stock analysis within the indoor and environmental quality (IEQ) topic shows a rapid increase with the focus on comfort, indoor air quality, and natural lighting. Supervised and reinforcement learning was mainly exploited for monitoring or customization purposes. Studies regarding indoor comfort could be, for example, optimal control of active and passive ventilation systems. Chen et al. (38) developed optimal control decision models using reinforcement learning for HVAC and window systems to minimize energy consumption and thermal discomfort. The reinforcement learning control system assessed the outdoor and indoor environments and was more effective and cost-efficient than heuristic control. Traditionally, indoor air quality (IAQ) is evaluated by measured data in mechanistic models. However, it has limitations in reflecting the occupied environments and process measurements at a large scale. Because of the shortcomings, machine learning and statistical models are getting popular to study indoor particulate matter concentrations, carbon dioxide, and radon (19). Among all techniques, artificial neural networks, multiple linear regression, partial least squares, and decision trees are classifiers that frequently study IAQ parameters. Kropat et al. (39) employed ensemble regression trees to map and predict multidimensional influences on indoor radon concentrations. The results of k-medoids clustering of lithological sample units also help to interpret radon properties of rock types. Besides, Seo et al. (40) proposed an occupant-oriented indoor luminous recommendation system with the help of the machine learning algorithm. With the data input from lifelog data and luminous environment data, a customized luminous environment was recommended according to task type, fatigue level, and emotion class.

Machine learning applications in the climate topic involve the impact of the built environment on the natural resources or environment, such as water consumption, renewable energy load, and carbon dioxide emission. The first part, CO_2 emission, is exemplified by a study by Javanmard et al. (41). The integrated energy-water optimization model in buildings was constructed for machine learning algorithms to predict costs and carbon dioxide emission. The investigating model conditions attained high prediction accuracy in various geographical regions. Additionally, Leo et al. (42) investigated the performance of a machine learning-based multi-objective prediction framework for multiple building energy loads. Their outcome indicated that artificial neural network facilitates effective building energy management by predicting accurate demand of heating, cooling, lighting loads, and building integrated photovoltaic electric power production.

To realize the development of smart districts, integration of renewable energy sources, loads, and electric vehicles (EV) is necessary. Longo et al. (43) presented the possibility of using the artificial neural network to forecast solar irradiation and load power consumption. Their results pave a step forward regarding optimizing electricity balance from renewable sources using EV batteries at the district level. Following the same context, machine learning was used for optimizing a photovoltaic(PV)-based energy system with battery and electric vehicles by Rehman (44). More specifically, algorithms were employed to identify overproduction curtailment and electric vehicle charging events to generate profiles of the PV energy production, feasible time windows, energy requirements for EV charging, and building's energy demand. The empirical results indicated the potential of achieving the net-zero energy balance through optimizing the combined energy systems. Finally, the performance of machine learning models for the implementation of demand response strategies was compared with the rule-based approach in smart-grid ready residential buildings. The study by Pallonetto et al. (45) showed that the predictive algorithm outperformed the rule-based approach to control an integrated heat pump and thermal storage system in terms of economic and environmental aspects.

The results from the extensive literature review show that applied machine learning approaches enable the building stock research to overcome several existing barriers. First of all, the diverse model types have the capability to process large amounts of heterogeneous data types. The analysis dimension can be expanded by coupling economic, societal, and environmental data to measure the building stock development with an overall picture. Next, the strong computational power of the models helps extend the geographic scope from a single case study to regional studies. Leveraging the statistical learning from the subset of observations, the performance of predictive algorithms exceeds the rule-based control without specifying complete assumptions. This ability to detect underlying patterns from data is useful when collecting data is resource demanding, which is also the natural limitation in building stock domain. As the BuiltHub datasets involve various topics, NUTS levels, and time-frequency, identifying the complete datasets to formulate relevant hypotheses will be the first step before data inquiry. Building profile clustering and future demand prognosis could be a potential direction in modeling considering the BuiltHub dataset features. Hence, feasible application of cluster techniques, tree-ensembled methods as well as neural networks should be further studied.

Using machine learning as part of the BuiltHub roadmap

BuiltHub Deliverable D6.6 details the need for open data and the various benefits Member States can have in the Political, Economic, Social, Technological, Legal, Environmental and Implementation dimensions. This is done as part of the work to devise the BuiltHub roadmap that will be offered as an support to collaborators.

Building-specific data are fundamental for developing digital applications in building stock management. Such kinds of empirical data can be retrieve from building permits, inspection records or building measurements. To translate the unstructural data and merge them with building registers, standardization is necessary to assure the data quality i.e., reliability and compatibility. Accessibility to these generic and specific building data sources offers the opportunity to apply the data-driven approach in the building stock analysis and is promising to be replicated internationally. If the data mining of environmental data and the subsequent descriptive analysis prove beneficial to the C&D industry, then a similar method can be applied extensively to assess other in situ materials in the European building stock for prospective material recovery.

In this section, two examples on how building environmental data contributes to quality assessment and recycling of in situ materials are provided. Example I concerns *tracing and evaluating the recycling potential of PVC flooring* through “screening relevant data”. Then Example II presents “algorithm development” and “technical solution delivery” in the project of *developing machine learning-embedded applications to support hazardous material assessment in renovation and demolition.* Furthermore the roadmap also includes a business model for long term financial viability. Machine learning and analysis are tools that will be part of a service that can be provided as part of the business model. Central to this is the accessibility of data. If BuiltHub has unique accessibility to data and potential query from stakeholders, then the business case is stronger. This is illustrated in the Sweden Pilot case described in this deliverable.

Workflow development

To facilitate the building stock analysis using machine learning techniques, an applied machine learning loop was developed in the RISE Research Institutes of Sweden and applied to T4.4 in the BuiltHub project. The diagram of the loop presented in Figure 3 consists of five sequential steps: (1) organization needs, (2) domain consultation, (3) data screening, (4) algorithm development, and (5) technical solution delivery. Each step is driven by the outcome of the previous step and thus is indispensable to complete the iteration. The workflow for the BuiltHub building stock analysis will take reference from the loop and identify the necessary tasks to move forward.

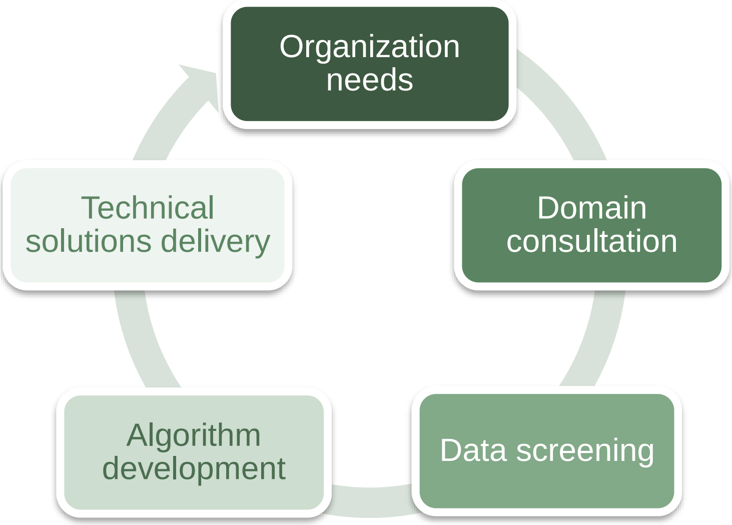


Figure 3. The applied machine learning loop.

Organization needs are the foundation for conducting building stock analysis that the problem owners define the scope of machine learning tasks. The stakeholders' interest is essential for navigating general assumption, testing and delineating data query boundary (see example I, II and III below). Therefore, a data-driven study concerns the match between the understanding of present limitations and the availability of relevant datasets. The outcome from Task 2.2 stakeholder interviews guided the focus of building stock analysis in uploading and framing the Swedish case on the BuiltHub webpage. In this way, a comprehensive perspective of the status quo, domain experts should be involved after identifying the organization's needs. These researchers and practitioners play a critical role in assessing the viability of the analysis and setting up the strategy for implementation. The feasible requirements for data acquisition, key variables connection, and influential feature selection will also be evaluated at this stage and followed by the step of data screening, which involves identifying potential datasets or field data relevant to conducting analysis. Within building stock research, limited open databases or digital datasets are available due to the lack of the tradition of digital documentation. Therefore, researchers tried to develop new tools, such as applications or online platforms, or exploit existing registered data, such as building registers or building permit documents, to assist the data collection process. From a long-term perspective, it would be beneficial to establish a generic digital building information database and, in addition, enrich the database with specific thematic information. In this way, data processing time, including dataset merging and cleaning, can be shortened since the datasets are structured, consistent, and machine-friendly. Data validation and transformation are required to be finalized to avoid the risk of generating biased results from a skewed dataset.

After controlling the dataset quality, the next step is to develop an algorithm according to the problem types. The work usually begins with explorative data analysis to understand the underlying data structure and the amounts of missing values. Histograms and boxplots are popular ways to visualize the parameters in descriptive statistics. Besides, the correlation of variables can be studied through pair plots, pairwise matrix, or multivariate regressions. Then based on the data type of dependent variables, prediction targets concerning regression (for continuous variables) or classification (for discrete variables) can be determined. Furthermore, choosing the machine learning models based on the criteria evaluated in Section 4.2. *Feature selection* concerning choosing critical predictive variables can be conducted with the help of domain knowledge or feature engineering algorithms. Through carrying out the stepwise exercises, the prediction performance of different machine learning models can be evaluated. After attaining the results, cross-validation work can be initiated to prevent the models are overfit or underfit to the training dataset. Please see the added examples I, II and III for a more detailed understanding and some practical cases.

If high accuracy rates of the models are verified, the next step is to interpret the output and deliver technical solutions. This work is highly associated with step 1 *Organization needs* and step 2 *Domain consultation* to transform the technical insights into an executable plan. The technical solutions can be delivered as an expert system combined with a decision tree or general suggestions, depending on whether the outcome will be used for property owners or policy measures. Two empirical research examples that adopted the applied machine learning loop will be presented in the subsequent sections to embody the loop concept in a practical term.

Three examples in the field of building stock analysis are selected to show how the applied machine learning loop can be incorpintoate into the research process. They are chosen due to their close association with the BuiltHub thematic areas and can reflect the potential analysis for the BuiltHub datasets. The first example relates the thematic area 2.1 Building stock characteristics, while the second example concerns 2.2 Building renovation. Through illustrating their study objectives, data input, and project outcomes can facilitate evaluating the potential obstacles when implementing the workflow to the BuiltHub project. Furthermore, Example I concerns the essential topic of hazardous material identification to realize a circular economy in the construction sector. Also, the results highlighted in the study can add the ongoing discussion on including the whole lifecycle in building stock analysis. The subject of Example II responds to the relevant EU policy for the BuiltHub project, including Energy Performance in Building Directive (EU) 2018/844 and Renovation Wave (COM) 2020/662. Thus, these two studies are valuable to be demonstrated in the following sub-sections.

Merging of datasets

Availability and quality of data is of the highest importance in AI and ML applications. The important source for building stock are the following:

- National registers, like energy performance certificate register and property register
- Local registers, like municipality building inventories and customer satisfaction surveys.
- Maps for instance building and property, demographic, land-use and infrastructure, and soil types.
- Height data, for instance, national height models.

The different data providers use different data granularity and ways of storing data. As a result building data are often stored in different data formats with obscure semantics and poor documentation which creates interoperability challenges when integrating and matching data for machine learning applications (46).

The accessibility and interoperability problems have prompted legislation such as the 2007/2/EC INSPIRE Directive to "create a spatial data infrastructure for the purposes of EU environmental policies and policies or activities which may have an impact on the environment" (I Directive 2007/2/Ec Of The European Parliament And Of The Council Of 14 March 2007 Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), 2007).

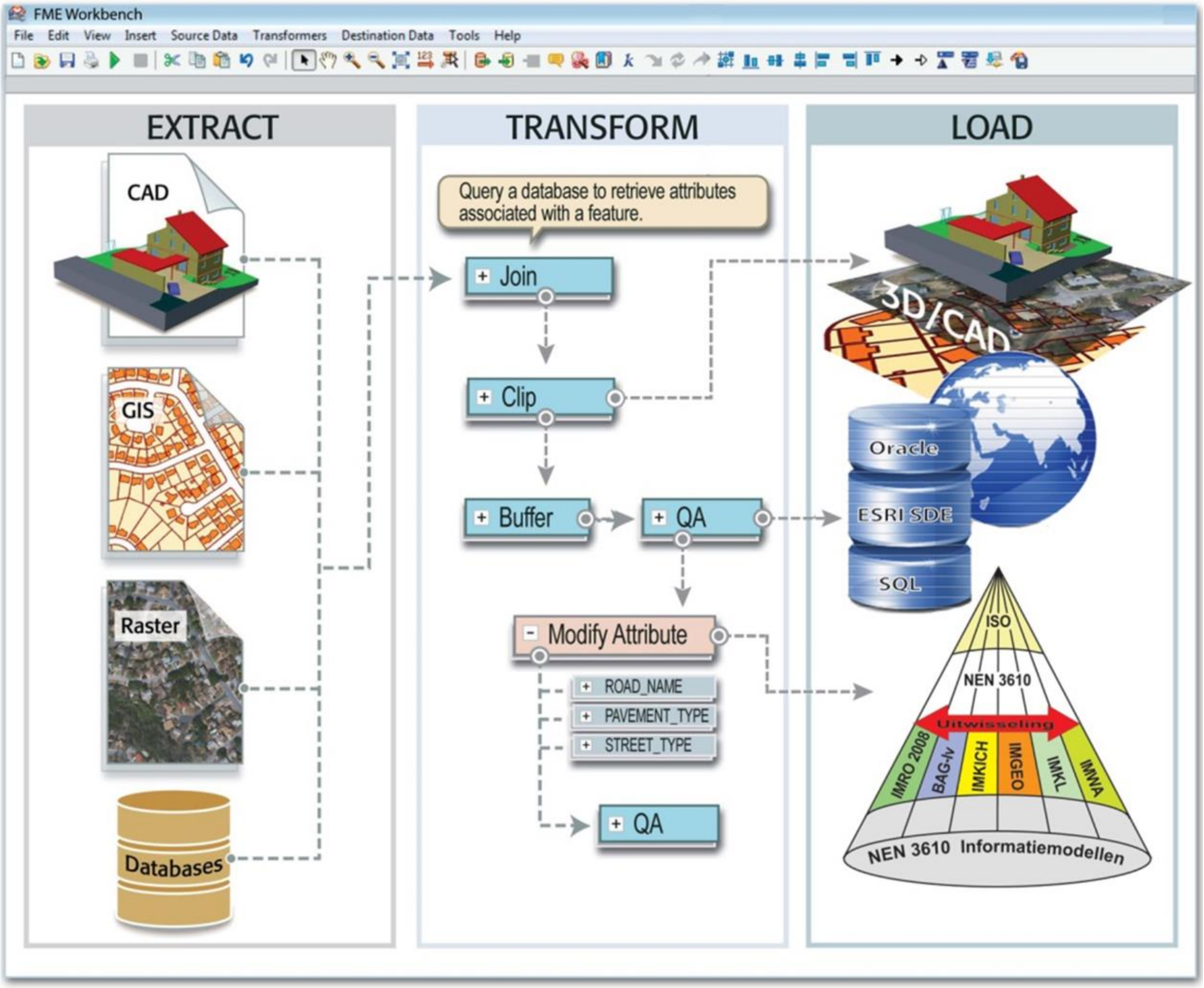
The Swedish mapping, cadastral and land registration authority, Lantmäteriet, has improved access to and the sharing of geo-data by creating geo-portals (e.g. www.geodata.se and maps.slu.se) and providing national specifications for geo-data to support (and improve the effectiveness of) national and local government services for urban planning, property registration, and management of building and environmental permits. These portals can bring additional value to e-government and support the growth of user communities and spatial services (47).

However, the challenges of merging dataset are still large and has not yet been overcome. Extract Transform Load technology was developed to support automated information integration from data warehouse. ETL processes collect data from different sources, then integrate and transform the data to support data analysis and decision-making within an organization (48). A data warehouse (DW) holds an integrated repository of information that is critical for business (49).

Recent years ETL technologies has also been used to prepare data for simulations and machine learning applications. The functions of ETL technologies are to extract, structure, aggregate operational data from databases and pre-compute queries to meet requests from users doing analyses (50–52).

Extracted datasets, of different formats and semantics, are subjected to a series of transformations to create homogeneous sets that can be stored in databases and delivered to end user applications (53).The transformation process involves several operations organized in a work flow, typically including schema transformations, cleansing, filtering, sorting, grouping and flow operations (routing and merging).

The ETL process can fix errors and correct for missing data, provide statistical measures, capture flows of transactions for safekeeping, combine data from multiple sources and structure data according to end-users' applications (52). An ETL workflow can be visualized using flow or network graphs (54), consisting of predefined readers and transformers connected to form the workflow in the ETL process (52,54). ETL tools used for implementation often have visual flow-based programming interfaces, which are easy to understand, allow re-use of parts of the workflow, and even enable non-programmers to create fairly complex programs with little training (55). Later research by Celani and Vaz (56) showed that visual programming interfaces accelerate the implementation of complex programs.



Query a database to retrieve attributes associated with a feature.

In previous studies by (57–59) the Feature Manipulation Engine (FME) is a spatial ETL that can handle both spatial and non-spatial data, as well as data from web services. It is based on the principle of “semantic mapping”, which enables the reconstruction of data during data conversion. Hence, FME can interconvert data more than 400 different spatial formats, including most of the common GIS, CAD and BIM formats (60). In addition, FME offers a variety of tools to perform spatial analysis, data exploration and geo-processing. The software includes approximately 450 transformers to perform different types of spatial and non-spatial operations (61).

ETL is being increasingly used to create performance models based on large quantities of spatial and non-spatial data. ETL tools are suitable for extracting data due to the interoperability possibilities they provide, for cleaning and conforming spatial data using diverse pre-defined functions (transformers), and finally loading the data back to file formats, databases or web services.

Analyses – missing information and clustering

In statistics, many techniques have been developed to fill missing data. The rule of thumb is dropping the missing data if the number of cases is less than the 5% threshold. Especially in the multivariate analysis, a large amount of missing data are usually dropped to prevent the risk of over-imputation and inaccurate inference of results (83). Two types of missing values should be distinguished for the applications of different handling methods. The first type is missing completely at random (MCAR), in which missing values are randomly distributed across all observations, whereas the second more common type is missing at random (MAR), in which missing values are distributed within one or more sub-samples (83). In MCAR, missing values can be handled by filling in the data with a t-test of mean different on the partitioning data or dropping the data with pairwise or a list-wise deletion (84). For missing data in MAR, filling the non-ignorable missing data can be achieved with multiple imputations such as maximum likelihood estimation, regression, using auxiliary variables to reduce bias, collecting follow-up data, and collecting data on intent to drop out (84). In practice, Python built-in library is capable of missing data detection, marking, and replacement. For example, filling in the missing values with a pre-defined constant value, referencing values from another randomly selected sample, using statistic features mean, median, and mode for the columns, and estimating interpolate values from a predictive model (85). In BuiltHub datasets, random missing data for a particular year may be interpolated from previous years. Yet, structural missing data above a certain threshold proportion should be dropped out.

Clustering is in the domain of unsupervised learning with sub-categories of K-means clustering and hierarchical clustering specific to time series data. Time series data has several characteristics that should be dealt with using particular clustering techniques, for instance, information in the ordered sequence, varied series length, and patterns that are not aligned in time between different series (86). Hence, combining the k-means clustering and the dynamic time warping algorithms can facilitate measuring the similarity between two temporal sequences. Time warping algorithm is a technique to calculate the optimal matching between two arrays. Firstly, clusters are constructed with k-means algorithms by splitting the data into k groups while minimizing the sum-of-squares in each cluster centroid, then employing dynamic time warping to collect time series of similar shapes (87). However, processing time-series data with k-means clustering can be slow for large datasets and thus gives rise to another alternative. Hierarchical clustering adopts a distance matrix to merge the least dissimilar clusters and visualize the clustering results in dendrogram (86). The advantage of hierarchical clustering is that the number of clusters does not need to be specified in advance; instead, adjusting a cutoff value results in different clusters. As the BuiltHub datasets contain high aggregated time series data with few observations, both methods can be tested in accordance with the proposed indicators.

Example I: Data-driven approach to trace and evaluate the recycling potential of PVC flooring in the building stock

Example I demonstrates implementation of the “data screening” step in the proposed machine learning workflow for specific building components – PVC flooring. The project aims to enhance plastic material recycling in the building sector by improvning knowledge on plastic use in the existing building stock. Currently, the knowledge of the types of in situ PVC flooring and its use in various building types are limited, making material recovery and quality assurance difficult. Characterizing the presence of PVC flooring and recycling potential in buildings can facilitate property owners and demolition contractors in their recycling work of old PVC flooring from demolished and renovated buildings. In fact, the entire flooring production value chain can benefit from more available and cost-efficient reclaimed material for plastic products, meanwhile, less non-recyclable wastes from renovation, deconstruction, and demolition for incineration.

Recycling plastic waste from renovated or demolished buildings is a promising opportunity to attain the climate-neural goal for the building sector. According to the EU EEA report (No 18/2020), the building sector accounts for the most significant share (around 69%) of PVC products with the longest product lifetime in Europe (62). The plastic components made from fossil raw material generate a high carbon footprint in the production phase, as well as significant environmental impacts for solid waste handling from end-of-life buildings. Considering the negative environmental impacts and future potential taxation on fossil-based plastic, increasing the extent of reusing and recycling plastic-containing components becomes an inevitable step to close the loop (62). Changing the business-as-usual linear value chain of plastic is also economically efficient to compensate for the growing need for fossil raw materials in production and reduce the risk of short of supply in new construction globally.

Among plastic materials in buildings, PVC flooring attributes a large number of plastic wastes in the construction and demolition waste (C&DW) compared to other plastic products, such as window frames, pipes, cables, and packaging (63). In the 90s, PVC flooring accounted for 51% of total flooring sales in Sweden, showing its extensive use in the past construction (63). Although a national system for separate collection and recycling of material residue from PVC flooring installation was established by the Swedish Flooring Association, the annual plastic residue recovery rate was less than 20% in Sweden in 2018 (64). A majority of PVC flooring collected as combustible waste was incinerated for energy recovery, which is the lowest level of waste hierarchy and resulted in an extra two tons CO2-eq per ton of PVC flooring recycled in Sweden (64). A recent LCA study on PVC flooring showed that bathrooms with PVC flooring as a surface layer have a higher environmental impact than bathrooms with ceramics tiles flooring (65). For example, the bathroom with PVC flooring contributes to an effect of 38 kg CO2 e/m2, the outer wall 18 kg CO2 e/m2, and the inner wall 11,5 kg CO2 e/m2 (65). To address the issue of limited recycling of PVC flooring, improvements on waste sorting of flooring by waste handling companies and extending producer responsibility principle to general flooring manufacturers are needed.

Old plastic floors represent a considerable material resource. Nearly 150 million square meters of PVC flooring were installed in the Swedish building stock, corresponding to over 350,000 tonnes of potential raw material with a recycling potential, which is equal to approximately one million tonnes CO2 (64). However, the information on the presence and the extent of PVC flooring in the existing building stock lacks systematic investigation. The building-specific information on flooring materials are not registered in the current national building database, making it hard to estimate the location and the extent of recyclable PVC flooring prior to renovation, deconstruction, and demolition works. Other barriers to low collection rates of PVC flooring are also identified, including a disconnecting recycling practice by individual manufacturers, a low adoption rate of the collection and recycling system in the sector, etc (64). To overcome these challenges, new knowledge on the historical use of PVC flooring in various building types in Sweden should be developed and transferred among the actors along the PVC flooring life cycle. Pre-demolition audit inventories from the renovated, deconstructed, and demolished buildings contain information on the presence of PVC flooring. Environmental investigations prior to demolition and extensive renovation are mandatory in many EU countries, regulated by local building codes or planning and building act (66). Over the years, the registers have been maintained by city archives in individual municipalities. By coupling PVC flooring records from pre-demolition audit inventories and building registers, it is possible to characterize PVC flooring in various building types in different regions built in the past century. This information is critical to the Swedish Flooring Association and C&DW actors to be better prepared to collect, separate, and recycle PVC flooring. Meanwhile, it can also be advantageous for property owners or demolition companies to estimate material assets in advance to plan for cost-efficient material recovery.

By compiling pre-demolition audit inventories from the three largest cities of Stockholm, Gothenburg, and Malmo, the environmental information, including plastic flooring used in a majority of building stock in Sweden, will be covered. The usefulness of such empirical data and the data integration workflow have been proved in the Swedish context (31). From an urban mining perspective, the gained information can benefit infrastructure and production systems at the sectoral level by promoting secondary plastic sourcing from the building stock. New knowledge of potentially available plastic flooring in the building stock can also encourage PVC flooring producers to transform their business models and value chains toward circular development with the scientific basis for investment decisions.

Three primary risks are identified to succeed in the project, described in Table 4. The first risk concerns available data on the PVC flooring from environmental inventories for each building type. Since plastic materials/wastes are investigated according to the resource and waste guidelines for construction and demolition, the risk of lacking sufficient data is relatively low. The second risk pertains to the quality of documentation on PVC flooring. The presence of PVC flooring in specific buildings and spaces are available, yet the availability of more detailed information on amounts and types of PVC flooring requires further exploration. Lastly, the timeline on various kinds of PVC flooring production and respective properties needs to be constructed.

Table 4. Risk identification and evaluation in the project.

Risk	Description	Likelihood	Impact
1	Data availability of environmental inventories, i.e., sufficient data amount for each building type to do statistical analysis.	low	middle
2	Data quality of environmental inventories, i.e., comprehensive information on types of plastic components and amount.	low	middle
3	Data accessibility of production records, i.e., timeline on historical use of PVC flooring types and their chemical properties.	low	low

Establishing a PVC flooring dataset in the building stock is expected to contribute to improving the existing national system for separate collection and recycling of PVC flooring in the built environment. For instance, *efficient sorting of plastic from the construction industry* can be achieved by disseminating the results along the C&DW value chain and the industrial-wide association. Next, the analysis outcomes from the empirical data can be used to streamline the guideline "Resource and waste guidelines for construction and demolition" from the Construction Federation (67). Promoting *efficient use of material resources in non-toxic circular cycles* is possible with the data-driven method as the information on asbestos-containing flooring and glue and PCB-containing acrylic flooring are available on environmental inventories to differentiate from clean PVC flooring. Lastly, the project outcomes can facilitate a better understanding of the available plastic materials assets in the building stock to *support other flooring manufacturers to assess secondary material supply and move toward green product portfolios*.

The plastic dataset is used as input data for statistical operation to forecast the recycling potential of PVC flooring in building stocks of other areas. The pre-study will form a basis for knowledge exploitation from practical sides. Building such expertise and research capacity is fundamental to ensuring the quality of the recyclable PVC flooring and improving its residue collection and separation rates. Primary barriers to PVC flooring separate collection and recycling, including engagement from manufacturers and standard practice within the industry, can be addressed through providing more transparent quality assurance of recycled PVC flooring. At the end, the system analysis results will be transform to technical recommendations for the PVC flooring waste management plan.

Example II: Development of machine learning-embedded applications to support hazardous material assessment in renovation and demolition

Example II shows how machine learning prediction can be used as a decision support tool in assessing the risk of hazardous materials in renovation and demolition projects. The project involve the areas of “algorithm developement” and “technical solutions delivery”. Over the years, the presence of hazardous materials in the ongoing rebuilding or demolition process posed high risks of health exposure to the demolition contractors (68), as well as 20% of cost increase and project delay due to acute decontamination for property owners (69). Identifying the potential presence of hazardous materials such as asbestos and PCB in advance can facilitate semi-selective demolition, a practice that removes contaminants or the material fractions that can overly reduce the quality of recycled building parts is regarded as a feasible strategy for implementing predictive maintenance (70).

On the other hand, the presence of hazardous materials also hinders circular economy-inspired actions in the building sector that is fundamental for reducing material-related carbon emissions during the production, use, and disposal phases (71,72). By recovering the materials and components’ values from the existing building stock and implementing circular economy-related strategies, 39% of global greenhouse gas emissions from material extractions, processing, and manufacturing of construction products can be prevented and 28% of virgin resource use can be cut (73). Despite the ambition toward a clean material cycle is unambiguous, several underlying factors prevent the construction and demolition waste (C&DW) sector from accelerating the uptake of reuse or recycled aggregates, above all, low-quality assurance of recycled materials, lack of control standards and tools, the needs for multiple processing before reuse, low market incentives for secondary materials, incomplete legal requirements, and so on (74).

To address the safety and sustainability aspects of building components, the EU Construction and Demolition Waste Management Protocol (75) and EU Waste Audit Guideline (76) were established to lead inventories of hazardous waste during the pre-demolition audit to ensure high-grade recycling and adequate waste management (75). Aligning with the legislative requirement, the implementation of hazardous waste inventory is mandatory in many member states for demolition, rebuilding, or extensive renovation permit application. Other regulations concerning hazardous material management, including handling and decontamination, are stated in national legislation such as Waste Regulations, Environmental Protection Agency regulation, and specific requirements from the municipalities. The accumulated environmental data are often maintained by individual municipalities and stored as hard copies in city archives (77). Owing to resource-demanding and time-consuming data collection and processing, the inventory data are left out from the national building database, which in turns hampers the understanding of reginal and national hazardous material stock (31).

The use of machine learning and data mining in building stock research has become increasingly common. Previous studies have tried to characterize the potential presence of asbestos-containing materials and leverage the key building characteristics to predict the likely contaminated buildings (69). For instance, a few machine learning applications have been developed in Poland to predict the spatial distribution and estimate the amount of the remaining asbestos cement roofing using hyperspectral images and national registers as input data (78). However, the majority of scientific publications for hazardous materials prediction at the building level are relatively few and remain at the statistical analysis.

As a starting point, the inventory data from Gothenburg and Stockholm cities were collected and coupled with national building registers to explore the possibility for predictive analysis, show in Figure 4. The pioneering efforts of creating a hazardous material dataset and hazardous waste inventory database have addressed the circular material aspect (31), which is left out from the interdisciplinary research domains of applied AI, C&DW management, and building stock analysis. Through improving the hazardous waste identification, source separation, and collection, the trust in the quality of recycled components/materials and the confidence in the corresponding waste management process will be enhanced. In the early stage of method development, detection patterns in specific building classes were confirmed and required to be verified with more input data. At the point of innovation implementation and knowledge exchange, the use of data-driven methods would need to be incorporated into organization needs and decision-making processes to deliver suitable applications for stakeholders.

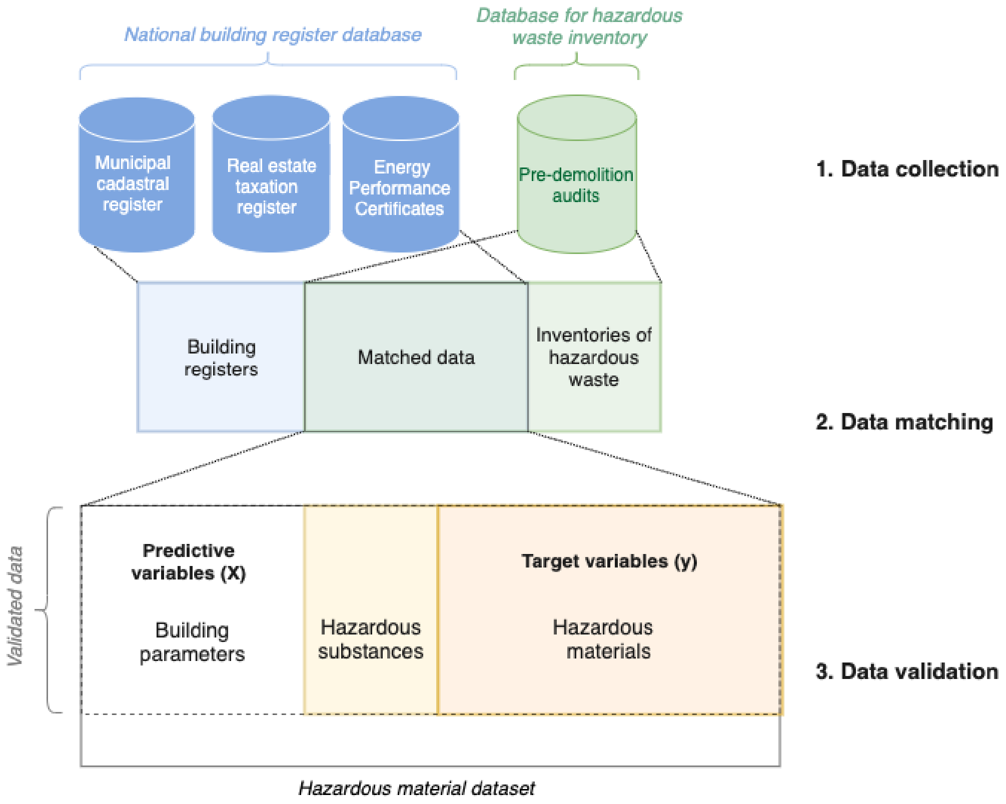


Figure 4. A proposed procedure for creating a hazardous material dataset by integrating and validating several data sources. Building parameters are in this case, predctive variables in modeling for binary classification of target hazardous materials.

In this deliverable, the authors developed a novel approach in the Swedish contexts by creating a data analysis workflow (Figure 4) and a machine learning pipeline (Figure 5) based on hazardous waste inventories to identify detection patterns of asbestos pipe insulation in multifamily houses and PCB joints or sealants and school buildings. In the explorative study, the prediction performance of various supervised learning classifiers, the minimum training data size, and the feature impacts and magnitudes to the model output were investigated (79). The attempted predictions showed method applicability for large-scaled risk screening and could be used to complement current environmental investigations.

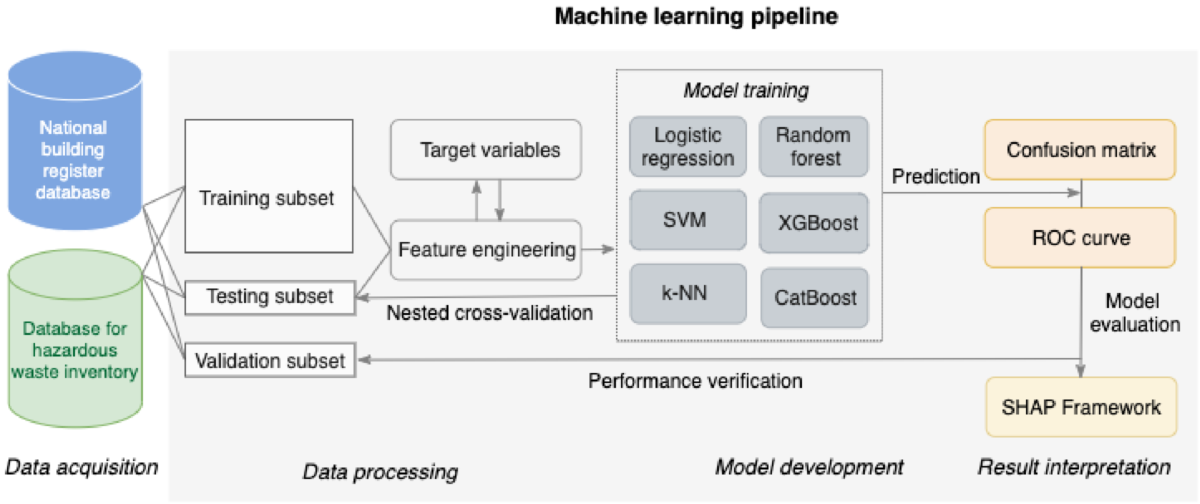


Figure 5. An machine pipeline comprised of (1) data processing concerning dataset partition and feature engineering, (2) model development including training and evaluation, and (3) result interpretation of influential features.

The previous work corresponds to the steps of domain consultation and data screening, and parts of algorithm development in the applied machine learning loop. The loop is developed by RISE as a framework to guide data inquiry for pre-defined hypotheses. Due to the outbreak of the COVID-19 pandemic, the research commenced with domain consultation to screen relevant field data and compile them into a machine-readable dataset. After that, machine learning models have been developed to identify algorithms with high performance for the specific task. At this stage, the quality and application of the inventory data have been confirmed, and the presence patterns of certain asbestos and PCB-containing materials in specific building classes have been ascertained. The next step is to engage the organization needs to clarify how the method can effectively contribute to the challenges in practice. The inputs from the stakeholders, including relevant authorities, property owners, auditors, waste handling companies, etc., are significant to delivering technical solutions.

The anticipated risk assessment decision support tools can help underpin the quality assurance of C&DW in the circular economy value chain and address the needs of actors in the value chain: (1) estimate the clean building stock on the macro-scale for the Housing, Building and Planning authority, and (2) predict the presence likelihood of hazardous materials in buildings on the micro-scale for demolition contractors and waste handling companies. Such decision support tools for hazardous material risk evaluation are not available in Sweden or other countries. Considering the asbestos and PCB hazard worldwide, the novel method development can establish a research front that can be of use for other countries. By puzzling up the last three elements of the applied machine learning loop, the research applications can promote the ongoing progress toward the circular and contaminant-free built environment.

To summarize, the aim of the project is built on the previous research results and further exploits the application potential of machine learning models in response to the needs of the public and private sectors. The project outcomes are expected to contribute to a closed building material loop and the proposed process and service innovation have a high replicable potential internationally. For instance, the EU Commission recently published a working document on

“Scenarios for a transition pathway for a resilient, greener and more digital construction ecosystem” (Brussels, 14/12/2021), in which the importance of developing a digital hazardous material inventory or building logbook is stressed. As such, the project results — machine learning-embedded hazardous material assessment and tailored deconstruction planning strategy — can address this emerging area and help EU countries create a more digital and circular construction ecosystem.

Example III: Development of machine learning-embedded applications to predict cultural heritage values

This example is a recent development of an analysis with the purpose of predicting cultural heritage values in the Swedeish building stock. EU requires member states to make lists of all buildings that are of special cultural heritage value in order to preserve cultural heritage values while still improving overall energy efficiency (in the EPBD). In Sweden no such register exists. Instead there are multiple registers in which building have been added one at a time over the years. There is thus a risk that some building that sould be on some of the lists have been overlooked.

In this example machine learning was used to predict cultural heritage value based on the existing lists and google stree view images of all buildings in Sweden. The work flow process can be seen in figure 6.

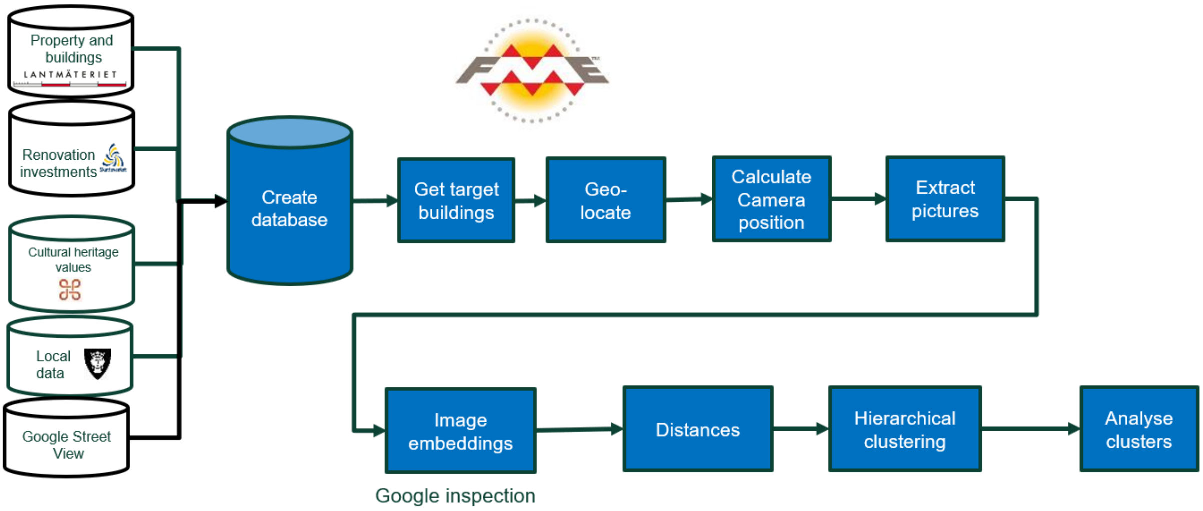


Figure 6 Workflow for prediciting cultural heritage values

The final output were clusters of buildings which were used by cultural heritage experts to speed up the process of creating the complete Swedish list of buildings with cultural heritage value..In figre 7 below examples of clustered buildings can be seen.

1570.html

1577.html

3.html

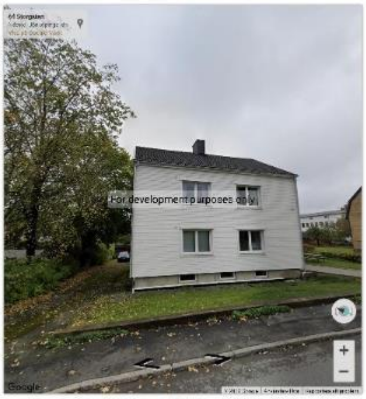
455.html

1219.html

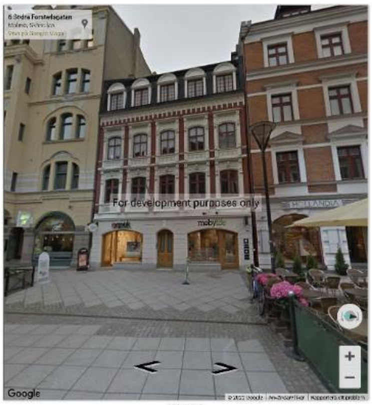
332.html

577.html

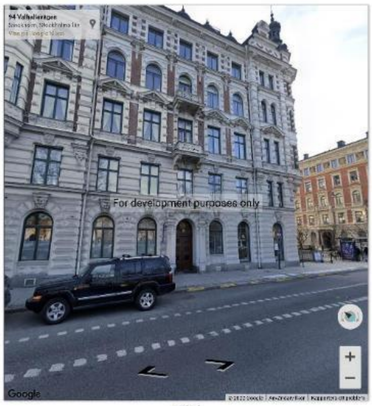
883.html



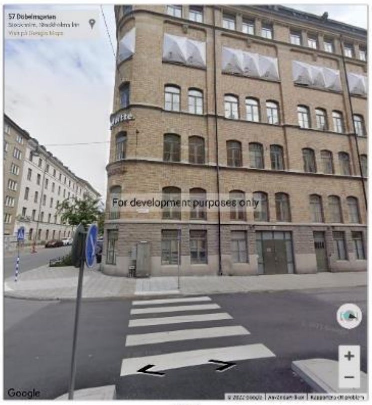
918.html



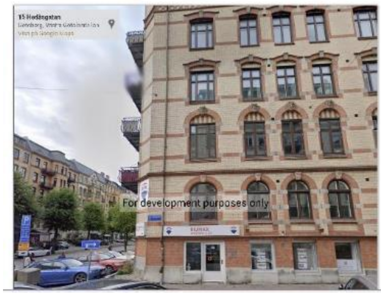
1254.html



261.html

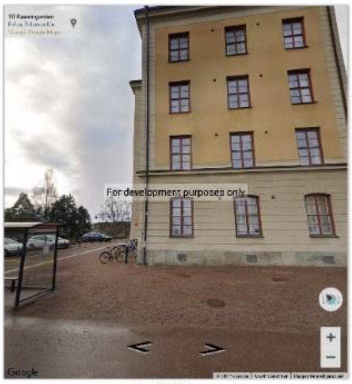


360.html

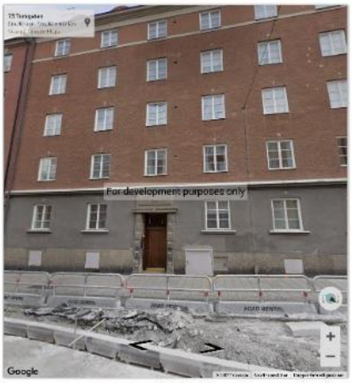


1234.html

46.html



693.html



705.html

964.html

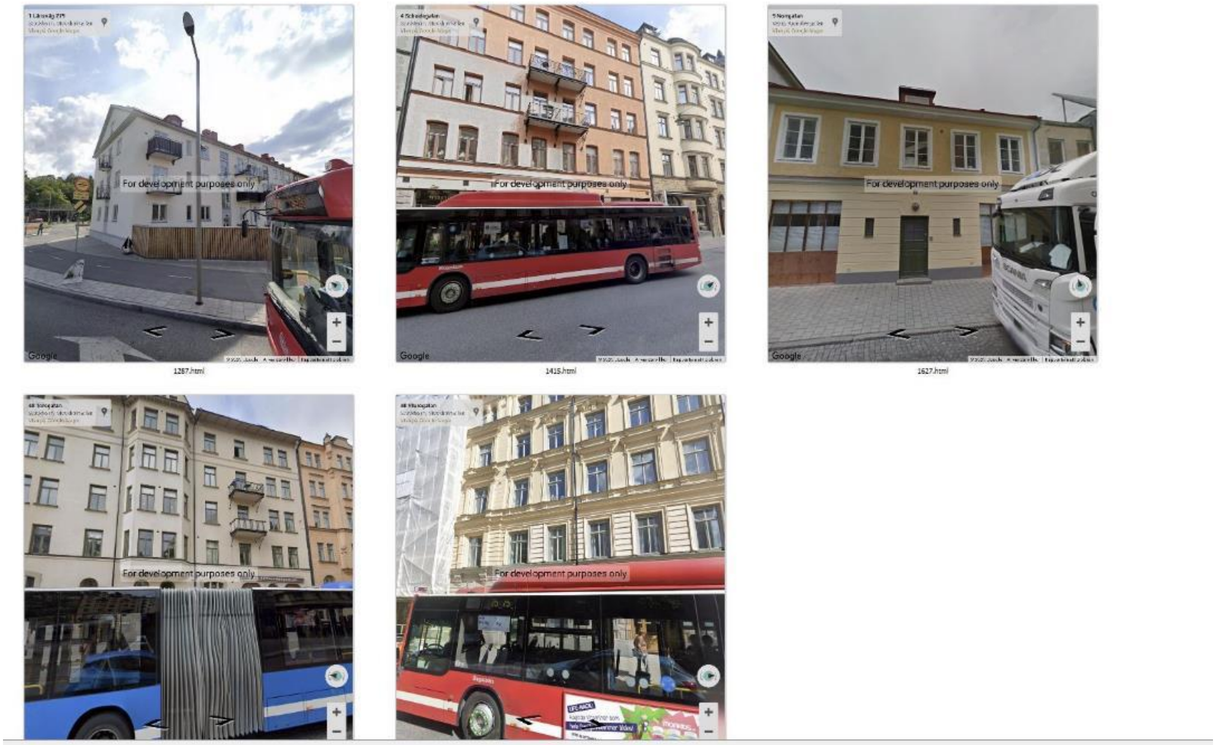

For development purposes only
For development purposes only
For development purposes only
For development purposes only
For development purposes only
Google
Google
Google
1287.html
1415.html
1627.html

Figure 7 a-e, Different building image clusters that were automatically generated from google street view images. Figure 7e illustrates a methodological problem. Images that contain busses have been clustered into one separate group.

Example IV: Usage of fuzzy logic to ascertain building owners

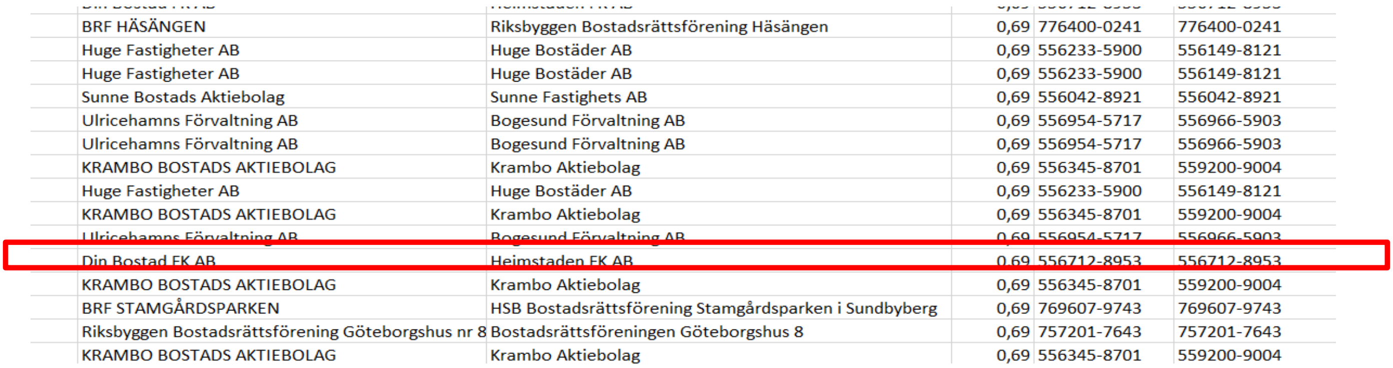
Comparing registers with property ownership to identify changed ownership over time can be difficult due to the old organization numbers remains unchanged and hierarchical company structures. Small changes in organization names can further complicate the identification of ownership changes. For this reasons fuzzy logic can be used for the improvement of data. That enables:

1. Utilization of information, it is possible to study the impact of rental housing acquisitions on the largest rental developers.
2. Understanding the transitions in ownership can help us to understand the transformation of the building stock and support decision making

In figure 8 two company names of building owners are compared the fuzzy logic score is seen in the third column. In this example we used the following steps to assure comparability between the building owners:

1. Two company names were compared to identify any differences in spelling or syntax using a fuzzy matching algorithm that generated a score based on the level of similarity between the names.
2. If the company names were an exact match, or if the fuzzy score was very high, this suggested that no change in building ownership had taken place.

3. If there were slight variations in the names, or if the fuzzy score was lower, further investigation was conducted to determine whether a change in building ownership had indeed occurred, despite the retention of the same identification number.



Din Bostad FK AB	Heimstaden FK AB	0,69	556712-8953	556712-8953
BRF HÄSÄNGEN	Riksbyggen Bostadsrättsförening Häsängen	0,69	776400-0241	776400-0241
Huge Fastigheter AB	Huge Bostäder AB	0,69	556233-5900	556149-8121
Huge Fastigheter AB	Huge Bostäder AB	0,69	556233-5900	556149-8121
Sunne Bostads Aktiebolag	Sunne Fastighets AB	0,69	556042-8921	556042-8921
Ulricehamns Förvaltning AB	Bogesund Förvaltning AB	0,69	556954-5717	556966-5903
Ulricehamns Förvaltning AB	Bogesund Förvaltning AB	0,69	556954-5717	556966-5903
KRAMBO BOSTADS AKTIEBOLAG	Krambo Aktiebolag	0,69	556345-8701	559200-9004
Huge Fastigheter AB	Huge Bostäder AB	0,69	556233-5900	556149-8121
KRAMBO BOSTADS AKTIEBOLAG	Krambo Aktiebolag	0,69	556345-8701	559200-9004
Ulricehamns Förvaltning AB	Bogesund Förvaltning AB	0,69	556954-5717	556966-5903
Din Bostad FK AB	Heimstaden FK AB	0,69	556712-8953	556712-8953
KRAMBO BOSTADS AKTIEBOLAG	Krambo Aktiebolag	0,69	556345-8701	559200-9004
BRF STAMGÅRDSPARKEN	HSB Bostadsrättsförening Stamgårdsparken i Sundbyberg	0,69	769607-9743	769607-9743
Riksbyggen Bostadsrättsförening Göteborgshus nr 8	Bostadsrättsföreningen Göteborgshus 8	0,69	757201-7643	757201-7643
KRAMBO BOSTADS AKTIEBOLAG	Krambo Aktiebolag	0,69	556345-8701	559200-9004

Figure 8 Etracted image as an example of how application of fuzzy logic algoritm output can look. The 0,69 value is the factor that can be used to separate the companies that are considered a match and those that do not match.

The Swedish pilot case: Using machine learning to enrich the building database for energy retrofitting

The Swedish pilot case demonstrates the implementation of machine learning to complement additional information to the building database for planning energy retrofitting from a study by Platten et al. (32). The dataset used to train the ML models, a sample of the full national building-specific dataset, metadata, and the calculation behind the key reference in the work (80) were provided to the BuiltHub consortium as part of the RISE commitment, on the 16th of June 2022, and is also attached to the D4.4 submission. This data is the input data used for the ML models. The results are presented in this chapeter.

How analyses can be run on the platform, and what infrastructure would be needed. How it could be implemented still needs to be worked out with the BuiltHub partners that are working with the platform. At the current moment, RISE is still waiting for the reply of what is needed to take the Swedish pilot case further in this regard. However, the delay is probably due to the fact that the Flanders pilot case has taken a lot of time and is a higher priority as the Flanders case involves external partners that have ongoing needs for analyses. As opposed to the Swedish case which is based on previously conducted work.

Worth mentioning is that the Swedish pilot case also demonstrates how a business model can be derived from working with machine learning in the institute/academic sector. In Sweden, as well as in many other countries, there are different rules for data access for academia and the private sector. There are analyses of building stock data that the private sector cannot provide to authorities. Only research consultancy by an institute could meet the data access requirement. Universities in Sweden have a tendency to give research consultancy work a low priority since it leads to few and more applied scientific publications.

Building characteristics are essential information for investigating the feasibility of specific energy conservation measures. However, there is a lack of records to enable evidence-based national energy policy (organization needs). With respect to authorities' needs, a study to enrich the EPC database based on the prediction results of building characteristics was designed. The research scope was limited to multi-family houses built between 1945-1975, which face increasing renovation demands with potential energy retrofits (domain consultation). Screening the data source that can complement EPCs with building-specific information, including building type, façade material, and eaves overhang, was executed with the help of Google Street View (data screening). Through combining a limited number of expert observations and the generation of interpretable machine learning models, unknown building characteristics relevant for determining specific energy measures can be predicted.

For this study, a reference with tailored energy retrofitting packages for the Swedish multifamily building stock 1945–1975 have been used (see Reference (80)). These energy retrofitting packages will be used to exemplify how building database enrichment can enable more accurate national estimations of energy savings and costs, as a certain number of building characteristics are needed in order to allocate appropriate energy retrofitting packages to buildings. For each building type described in Section 2.1, there are three available packages (1–3) which entail di_erent costs and energy savings (low to high). The packages must be applied in successive order, meaning that Package 2 requires Package 1 to have been conducted, and Package 3 requires both Package 1 and Package 2 to have been conducted.

- In Package 1, a number of measures that aim at optimising the operation of the building are undertaken (80). Apart from building type, no building characteristics must be known in order to determine the feasibility of the measures in Package 1.
- In Package 2, components such as pumps and fans are changed to more e_ective counterparts, and additional insulation is added in the attic and to existing windows [44]. As for Package 1, building type is the only characteristic that needs to be known in order to determine the feasibility of the measures in Package 2.
- Package 3 contains the most extensive measures, including a new ventilation system with heat exchange from exhaust air, a change of windows, and 10 cm additional insulation on the building envelope (80). To determine the feasibility of Package 3, two building characteristics apart from building type are of advantage to know. The first characteristic is the façade material; more specifically, it is of advantage to know whether the building has a brick façade or not, as brick facades often must be preserved due to cultural and historical values. Additional insulation on a brick façade is thus not always a feasible option. More so, the shape of the roof and length of the eaves determines whether there is room for additional façade insulation or not, and additional façade insulation on buildings with an existing eaves overhang thus involves less extensive inventions than when the existing roof must be adjusted to a thicker façade. Consequently, eaves overhang is a necessary building characteristic to know to determine the feasibility of Package 3.

Ocular observations in Google Street View were conducted for 476 EPCs that were sampled from the total of 50,000 EPCs 1945–1975. The sampling was performed as weighted random sampling, where the probability of each EPC being selected was determined by the area of the building the EPC represented. The reason for the weighted random sampling was to gain a sample that was representative of the building stock in respect to area rather than in respect to the individual EPCs. However, due to a low representation of certain building types (tower blocks) in the sampled data, observations were conducted for another 41 manually selected EPCs, resulting in a total of 517 observations. The manual selection of complementing EPCs was based on number of storeys, as tower blocks usually are higher than slab blocks.

For the sample of 517 EPCs, observations were conducted in Google Street View using the registered address in the EPC. Observations were conducted by all of the authors, and methods to ensure that observations were conducted uniformly were undertaken. The quality of the observations was ensured by first letting all authors make observations guided by a senior researcher. After that, a control matrix of 13 observations was constructed to ensure that all authors' classifications agreed. After corrections, authors conducted observations individually. Any ambiguous cases were discussed with a senior researcher before classification, or rejection as valid observation. In three cases, observations were not possible due to lack of coverage in Google Street View. These EPCs were thus removed which resulted in a total of 514 observations. This was considered a su_cient number of observations as iterative testing of ML models starting at 200 observations showed no significant improvement in accuracy after 400 observations. The building characteristics and the respective classes that were observed are listed in Table 4. The choice of which building characteristics to observe was based on the gap between available data in the EPCs and data needed in order to assess the feasibility of energy retrofitting packages from the case presented in Section 2.3. It was found that the characteristics building type, whether or not the building has a brick façade, and whether or not the building has eaves overhang were needed. As seen in Table 4, rowhouses are included as a building type to be observed despite them not being introduced in Section 2.1. Rowhouses are not multifamily buildings per se, but the way they are owned determine whether their EPCs end up in the category for multifamily buildings or not. Rowhouses that are owned in similar ways as multifamily buildings (rental housing, resident-cooperatives) are classified as multifamily buildings in the EPC database, and they will thus be necessary to identify when using the EPCs to study the multifamily buildings 1945–1975. They will however be excluded from analyses of the energy savings potential in the multifamily building stock 1945–1975.

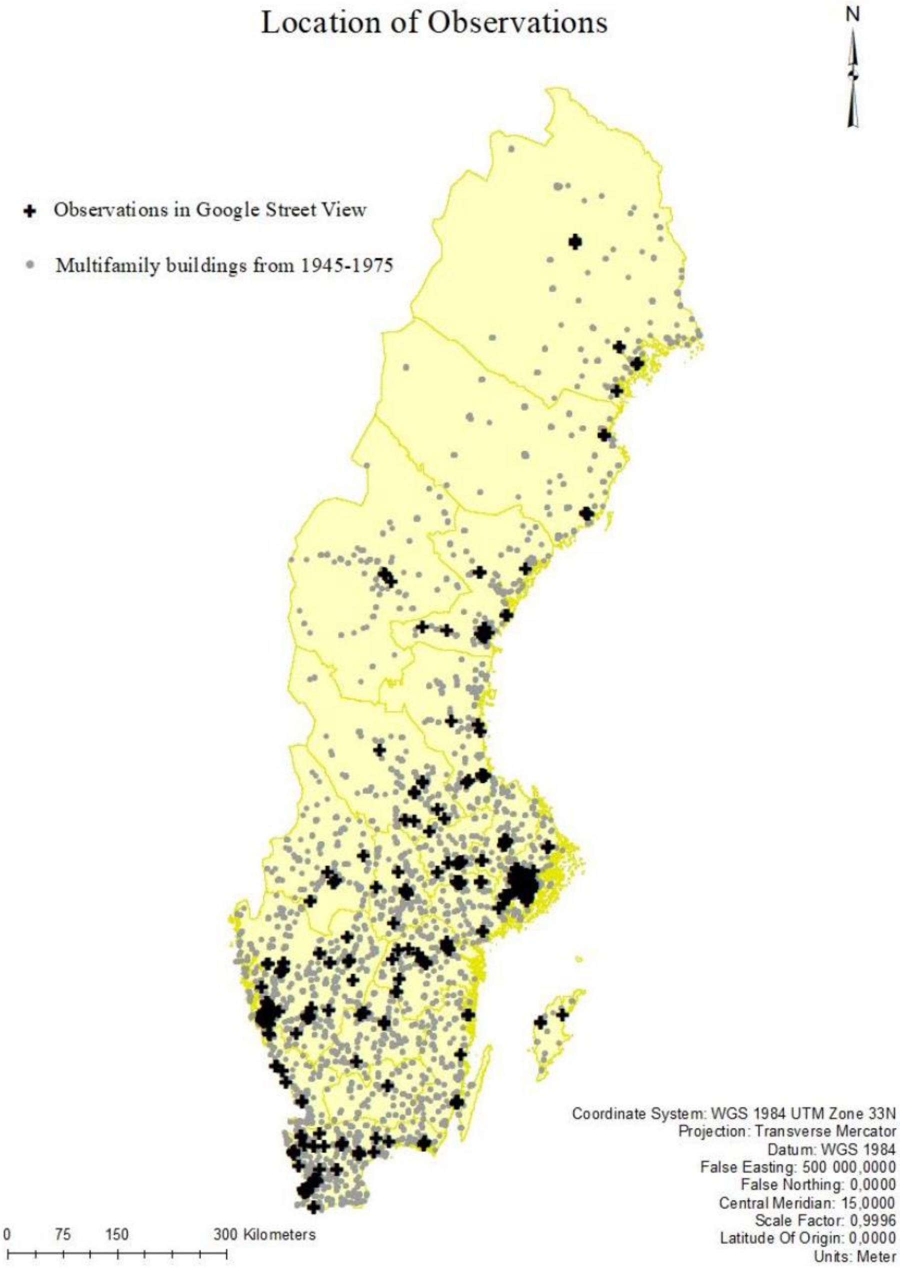

Location of Observations
N
✚ Observations in Google Street View
● Multifamily buildings from 1945-1975
Coordinate System: WGS 1984 UTM Zone 33N
Projection: Transverse Mercator
Datum: WGS 1984
False Easting: 500 000,0000
False Northing: 0,0000
Central Meridian: 15,0000
Scale Factor: 0,9996
Latitude Of Origin: 0,0000
Units: Meter
0
75
150
300 Kilometers

Figure 9. A map of Sweden showing the distribution of the multifamily building stock constructed between 1945 and 1975 (light dots) and the buildings that were observed in Google Street View (black crosses).

Finally, the geographical distribution of the 514 observations is shown in Figure 9. The light dots in the map show all multifamily buildings constructed between 1945 and 1975, whereas the black crosses mark the multifamily buildings that were observed in Google Street View. It can be seen that the studied multifamily building stock is distributed all across Sweden in a way that reflects the population density of the country. The observations show a similar pattern, indicating that they constitute a geographically representative sample.

The feature selection for machine learning modeling was executed in two parts: (1) pinpointing potential features by domain experts, (2) automatically validating the chosen features with stepwise regression. Derived features were generated from stepwise regression as ratios of two or more features to distinguish between building characteristics better. The number of stories and the construction year was found to be influential features to differentiate building types. Afterward, numerical variables were normalized with a min-max scale. Three supervised learning classifiers were explored for optimizing the prediction results (algorithm development). Considering the bias and variance trade-off of machine learning models, appropriate model types for the problem and careful parameter tuning and regularization were investigated. For instance, logistic regression has a high bias but low variance, whereas support vector machine has a low bias but high variance. 10-fold cross-validation was carried out in search of optimal machine learning model with the following criteria: (1) a high overall accuracy of training data with proximity to testing data, (2) distribution of accuracy for intended application purpose, (3) a low number of input features to enhance interpretability. Table 5 illustrates the chosen prediction model and its accuracy for each of the predicted building characteristics.

Table 5. The chosen prediction model and its accuracy for each of the predicted building characteristics adopted from Platten et al. (32).

Building Characteristic	Features in Selected Model	Machine Learning Model	Accuracy
Building type	Number of stories, Construction year Heated space per story and address Number of apartments per address	SVM	88.9
Eaves overhang + not brick façade	Construction year Number of apartments Number of stairwells per apartment Area code	SVM	72.5

As a result, the model that matches the criteria above was adopted to predict characteristics of the multifamily building stock of 1945–1975; thereafter, the predicted features were used to estimate energy-saving potential from various energy retrofit packages. The rapid construction of new buildings between 1945 and 1975 in Sweden was partly facilitated by standardized building methods and building types, which enables applying standardised methods for refurbishment and energy conservation measures for these different building types (81). This findings showed that almost all of the Swedish multifamily buildings built in this era can be categorized into four building types using the features in Table 4: slab blocks built before 1960, slab blocks built between 1960-1975, panel blocks, and tower blocks. These building types are fundamental characteristics of the building structure and correspond to different energy retrofit strategies concerning energy saving potential and cost estimation (80). Accordingly, a decision tree for tailored energy retrofit packages was developed based on four building characteristics in Figure 10: renovation status, EPC rating, suitability for additional façade insulation, and building type (technical solution delivery). These results can benefit the Swedish long-term renovation strategy for improving energy efficiency in the existing multifamily buildings.

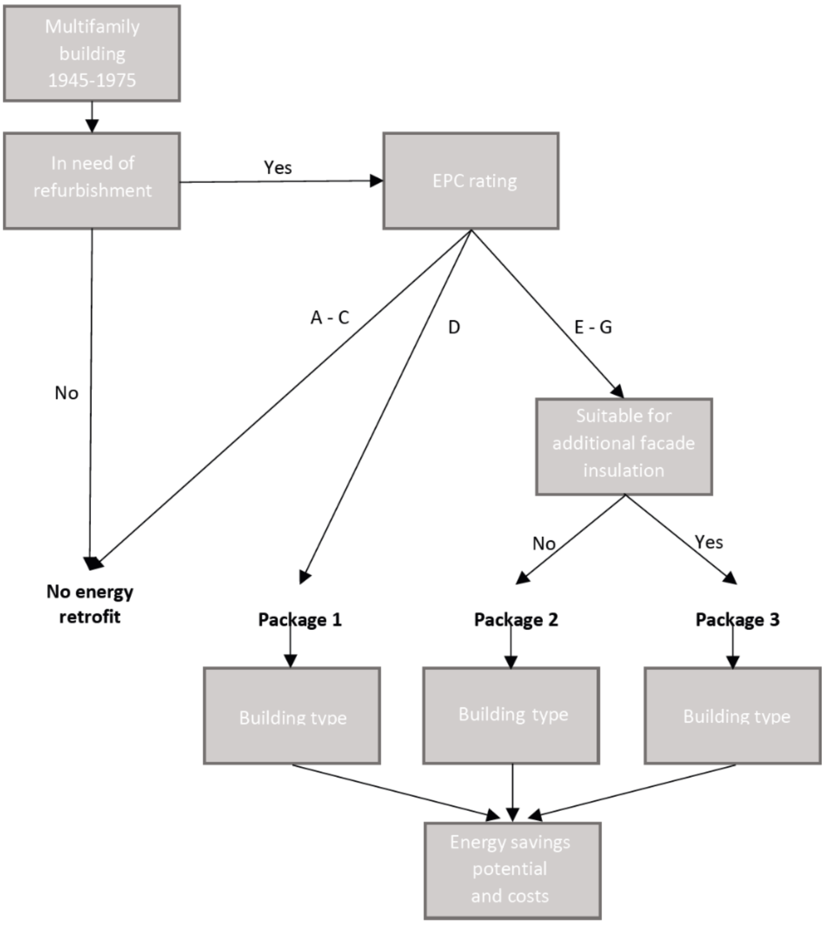


Figure 10. Decision tree showing how four building characteristics, i.e., renovation status, EPC rating, suitability for additional façade insulation, and building type, can help determine a tailored energy retrofitting package for each individual building adopted from Platten et al. (32)**.**

Results and application by the Swedish authorities

Using these methods, advice and basic data were provided to the Swedish authorities which were used in the formulation of the Swedish national strategy for energy efficncey. Figure 11 and 12 details the costs and resulting energy efficiency improvement that were presented in the final report tot he authorities.

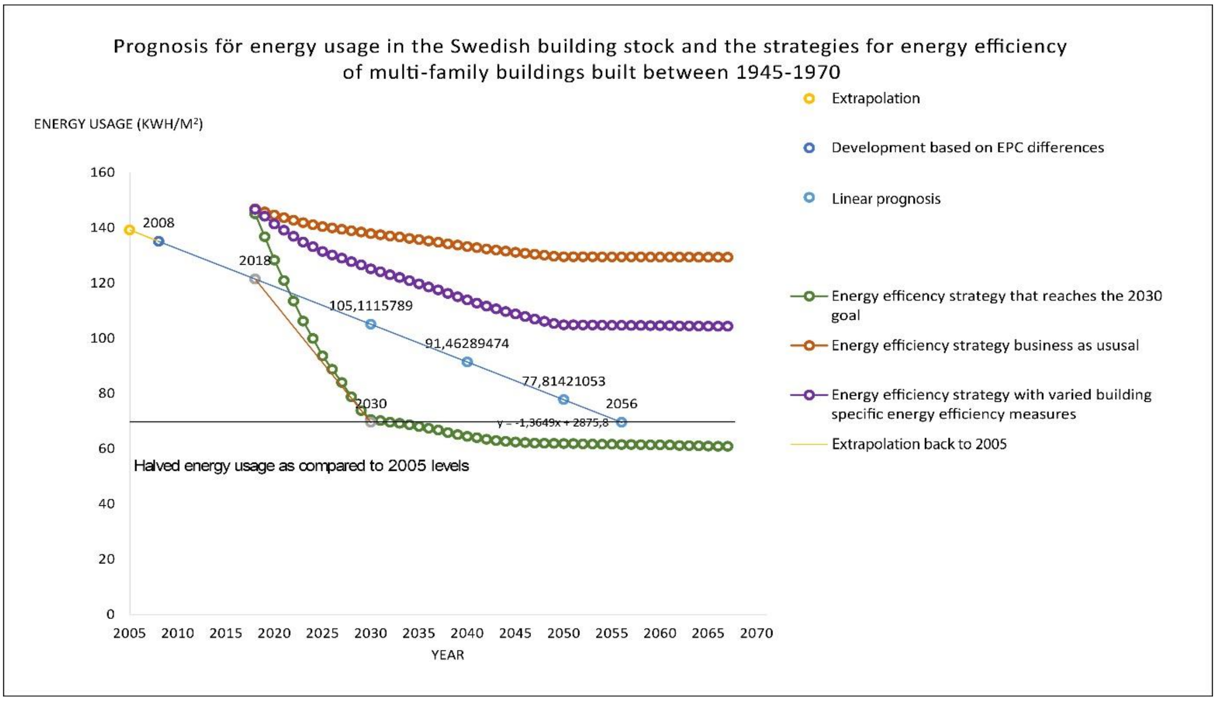


Figure 11 Energy usage. Extrapolated, previous, actual, prognoses and energy efficency strategies

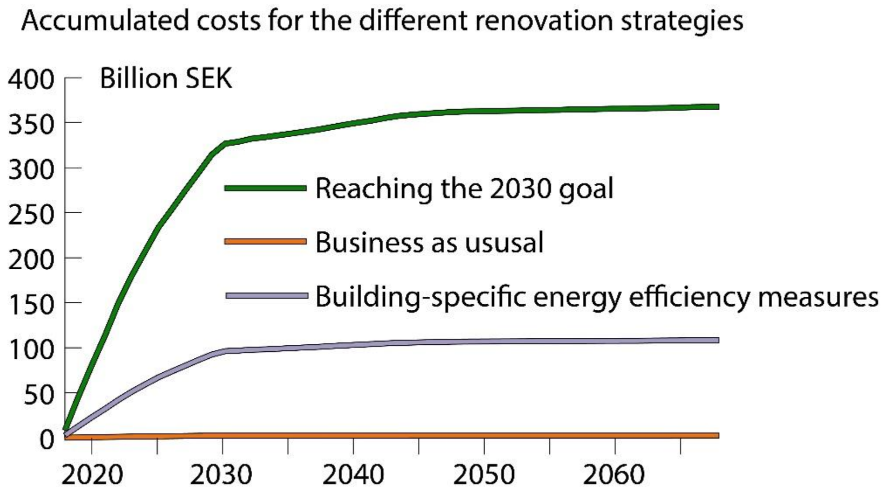


Figure 12 Costs for associated energy efficiency strategies

Based on the analyses the following concluding recommendations were made:

> *„The results of this report show that there is a large underlying need for renovation. This can be linked to the 'renovation debt' that exists and the necessary future investments that arise when the top of the 'million homes program' buildings have to be handled. These buildings will within the next few years have a value-year that exceeds 50 years. This great renovation need means that there are good opportunities to also make the stock more energy efficient. Energy efficiency strategies to reach the 2030*

target will lead to substantial costs. A particularly aggravating effect is also that the rate of new construction and that the need for housing is high, which also largely affects the renovation cost. This means that the construction cost with a higher percentage of total renovation would probably rise further. Population growth and the lack of housing further risk increasing the problems in the existing stock. And the renovation costs and cost for energy efficiency measures would increase the risks further.

Use of the value-year in calculations regarding renovations is currently the only method to estimate building-specifically the registered renovations and anticipated renovations. For example, in Mikael Mangold's PhD thesis, it was concluded that condominium associations often make several minor renovations and carry out more continuous maintenance than the other owner groups. Consequently, we can conclude that the largest investments have been made in the group of condominiums. These renovations have led to only minor energy efficiency improvements.

Concrete conclusions from the analysis of the apartment buildings built between 1945 and 1975 are that the assumed costs for energy efficiency should be reviewed. According to the assumptions, it seems important to quickly obtain permission for adjustments and operational optimization. Furthermore, it seems that slatted houses built before 1960 can form a larger group of apartment buildings that can undergo far-reaching energy efficiency improvements at a lower cost. Public utilities are overrepresented as owners of apartment buildings in this group.” (82)

The business case of using machine learning for analyses for authorities

The reworked version of the EPBD (to be presented the winter of 2023/2024) will require member states to have national building renovation plans. These plans should be linked with required political decision-making making and the plans should be designed to enable evaluation. Every member state will need to formulate and follow up these building renovation plans. This will be costly and cumbersome for many member states and there is an opportunity to help authorities in member states with these tasks using machine learning tools and analyses of building stock data.

The Swedish pilot case described in the previous chapter is an example of how this can be done. In the upcoming more extensive work of formulating national building renovation plans in Sweden, there is a plan to develop new machine learning tools for improved understanding of the renovation need in Sweden. Specifically, the new language models enable the parametrization of building data that is currently only available as text.

The new tools that the language models represent are novel and there are not many actors that have entered the market of selling building stock analyses that include data processing using language models. The combination of a larger upcoming need for building stock analyses among all member states and these new tools is a business opportunity that could be very suitable for BuiltHub.

The Swedish business case example

In Sweden, the most commonly used variable for estimating the renovation need is a variable developed by the Swedish Tax Agency to estimate buildings' need for renovation investments (Värdeår). When a renovation project is conducted that goes beyond maintenance it is registered by the Swedish Tax Agency. The cost of the renovation results in a change in the so-called value year of the building as described by the Swedish Tax Agency.

The purpose of recording a value year is to have an official record of the anticipated remaining service life of buildings. The value year is initially the year of construction but as renovation is conducted the value year will increase depending on the cost of the renovation as described in Table 6 and Equation 1. Registration of renovation in the tax index usually happens 1–2 years after the renovation.

Table 6. Methods for setting a *value year* based on renovation costs according to Swedish Tax Office.

	Renovation cost
Group 1	More than 70% of the *new building cost**
Group 2	20-70% of the *new building cost**
Group 3	Less than 20% of the *new building cost**

*The new building cost is increasing based on Inflation, changes in construction costs and property value. This is also specified in a table by the Swedish Tax Office.

$$\frac{Value\ year - Construction\ year\ [year]}{Renovation\ year - Construction\ year\ [year]} = \frac{Renovation\ cost\ \left[\frac{SEK}{m^2}\right]}{Cost\ of\ new\ building\ \left[\frac{SEK}{m^2}\right]}$$

Equation 1 How the value year is calculated for group 2 in table 6. For example: if a building built in 1960 was renovated in the year 2000 to a cost of 50% of the new building cost the value year after the renovation would be 1980.

The changes in value year only reflect the cost of the renovation but do not contain what kind of renovation measures were implemented. The value year is an indicator of renovation costs, or an indicator of investments into the building. However, the changes in value year contain the following uncertainties:

- More than one renovation can have happened, but only the last renovation year is registered
- It is not known what kind of renovation measure that has been conducted
- Property owners have different standards for what constitutes renovation and what is considered general maintenance.

Regardless of all these shortcomings, the value year is still the best proxy for estimating the renovation need in the Swedish building stock. RISE has previously mitigated shortcomings using statistical methods and cross-referencing with auxiliary data (specifically ownership data). RISE's idea is to make use of new data sources to further calibrate the analyses of the value years. The new sources are:

- Building permits
- Property owners' yearly audits
- Energy experts prescribed energy efficiency measures
- Real estate agent prospects

All these sources contain texts that can be analysed and parametrised using the language models. The benefit of using the language models is that they facilitate need-specific parameterization.

In the case of the requested national building renovation plans, there are several parameters that would be beneficial to map and take into consideration. First of all, to just calibrate the value year itself is one application. However, the type of renovations that have been conducted and the type of energy efficiency measures that have been suggested by the energy experts would be useful to better understand the renovation needs and the energy efficiency potential.

RISE has made initial tests with language model API and the Energy experts prescribed energy efficiency measures. The tests indicate that the idea is feasible in terms of parameterization quality and loading times. However, there is a need to purchase a licence and to work locally. Working locally is also a requirement to not make larger datasets available for the international language model providing companies.

Conclusions

A literature review on building stock research using machine learning provided a theoretical background for upcoming analysis. Overall, most of the building stock analyses were found in the topics of energy, building characteristics, and IEQ. The fact of uneven extent of research may be due to low data accessibility and high-cost factors. With emergent subjects in household EV charging and smart-grid ready buildings, more data can be expected to enrich the associated building stock analysis.

Also, a workflow for conducting an applied machine learning loop was proposed and exemplified. This agile process can guide the BuiltHub analysis by reviewing the presence of the necessary elements. Each step requires the BuiltHub working packages, and close communication is needed to deliver valuable solutions. These collaborative efforts can be regarded a: WP2 synthesizes the needs from stakeholders and consults domain experts for feasibility control of the hypothesis; WP3 compiles the data and performs data validation; WP4 conducts analysis and develops models; and lastly, WP5 presents the graph database and disseminates results.

Adopting the same concept, two more extensive research examples within the field of building stock analysis demonstrated the practical implementation of the loop in varied contexts. Predicting the presence of hazardous materials in buildings in Example I emphasized the importance of domain consultation and data screening. While using machine learning to enrich the building database for energy retrofitting in Example II reported the process of algorithm development and technical solution delivery. Both examples have a clear hypothesis for the prediction tasks and access to the validated, non-aggregated data. Despite the limited amount of labeled data, high data quality and data stratification process reduce prediction uncertainty risk. Following the context, filling the data gaps and conducting descriptive analysis were identified to be relevant in the subsequent tasks. Considering the BuiltHub indicators and existing dataset characteristics, the techniques for handling missing values and clustering for future work will be introduced in the next section.

Future work

The next step is to work with the Flanders pilot case. Various other BuiltHub WPs have started working and adapting data from the municipality. The intention is to use the Flanders pilot as a concrete BuiltHub example of the kind of support that can be given to a project partner stakeholder. The Swedish pilot case presented in this report will serve as a model for the upcoming work.

List of attached documents

Meta data on the Swedish EPC system before 2020 ‚Swedish EPC meta data 1.xls'

Meta data on the Swedish EPC system after 2020 ‚Swedish EPC meta data 2.xls'

The sample of the observations we made ‚Google street view observation sample.xls'

A sample of the resulting national register including final predictions ‚Sample of full sheet used for the national strategy.xls'

The reference we used to associate building types with building specific energy efficiency strategies and costs (SagaEllerVerklighet) ‚SagaEllerVerklighet_v2.xls'

List of tables

List of figures

References

1. Consortium B. BuiltHub | Home [Internet]. [cited 2021 Sep 8]. Available from: https://builthub.eu/

2. Kohler N, Hassler U. The building stock as a research object. Building Research and Information. 2002 Jul;30(4):226–36.

3. Kohler N, Steadman P, Hassler U. Building Research & Information Research on the building stock and its applications. 2010;

4. Pombo O, Rivela B, Neila J. The challenge of sustainable building renovation: Assessment of current criteria and future outlook. Vol. 123, Journal of Cleaner Production. Elsevier Ltd; 2016. p. 88–100.

5. Mangold M. Challenges of renovating the Gothenburg multi-family building stock. 2016.

6. Khayatian F, Sarto L, Dall'O' G. Application of neural networks for evaluating energy performance certificates of residential buildings. Energy and Buildings. 2016;125:45–54.

7. Leek J. The key word in "Data Science" is not Data, it is Science. Simply Statistics; 2013.

8. Dhar V. Data science and prediction. Communications of the ACM. 2013 Dec 1;56(12):64–73.

9. Data science - Wikipedia [Internet]. [cited 2021 Jun 28]. Available from: https://en.wikipedia.org/wiki/Data_science

10. Szokolay S V. Introduction to architectural science : the basis of sustainable design. Third. Abingdon, Oxon; 2014.

11. Nguyen AT, Reiter S, Rigo P. A review on simulation-based optimization methods applied to building performance analysis. Vol. 113, Applied Energy. Elsevier Ltd; 2014. p. 1043–58.

12. Hensen JLM, Lamberts R. Building performance simulation for design and operation. Abingdon, Oxon: Spon Press; 2011.

13. User interface - Wikipedia [Internet]. [cited 2021 Jun 28]. Available from: https://en.wikipedia.org/wiki/User_interface

14. Poole D, Mackworth A, Goebel R, Oxford NY, University O, Oxford P, et al. Computational Intelligence A Logical Approach. 1998.

15. Russell S, Norvig P. Artificial Intelligence: A Modern Approach, 4th US ed. [Internet]. [cited 2021 Jun 28]. Available from: http://aima.cs.berkeley.edu/

16. Darko A, Chan APC, Adabre MA, Edwards DJ, Hosseini MR, Ameyaw EE. Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities. Automation in Construction. 2020;112(January).

17. Hong T, Wang Z, Luo X, Zhang W. State-of-the-art on research and applications of machine learning in the building life cycle. Energy and Buildings. 2020;212:109831.

18. Wu H, Zuo J, Zillante G, Wang J, Yuan H. Status quo and future directions of construction and demolition waste research: A critical review. Journal of Cleaner Production. 2019 Dec 10;240:118163.

19. Wei W, Ramalho O, Malingre L, Sivanantham S, Little JC, Mandin C. Machine learning and statistical models for predicting indoor air quality. Indoor Air. 2019;29(5):704–26.

20. Mitchell T. Machine Learning. New York: McGraw Hill; 1997.

21. Koza JR, Bennett FH, Andre D, Keane MA. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In: Artificial Intelligence in Design '96. Springer Netherlands; 1996. p. 151–70.

22. Hu J, Niu H, Carrasco J, Lennox B, Arvin F. Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning. IEEE Transactions on Vehicular Technology. 2020 Dec 1;69(12):14413–23.

23. Conventional programming. Pcmag.com;

24. Expert system - Wikipedia [Internet]. [cited 2021 Jun 28]. Available from: https://en.wikipedia.org/wiki/Expert_system

25. KTH Computer applications in power systems - advance course. Introduction to Agent and Multiagent Systems.

26. Sutton RS, Barto AG. Reinforcement Learning: An Introduction Second edition, in progress.

27. Deep learning - Wikipedia [Internet]. [cited 2021 Jun 30]. Available from: https://en.wikipedia.org/wiki/Deep_learning

28. Deng L, Yu D, Deng L, Yu D. Deep Learning: Methods and Applications. Foundations and Trends R in Signal Processing. 2013;7:197–387.

29. Dalla Longa F, Sweerts B, van der Zwaan B. Exploring the complex origins of energy poverty in The Netherlands with machine learning. Energy Policy. 2021;156(September 2020):112373.

30. He Y, Henze J, Sick B. Continuous learning of deep neural networks to improve forecasts for regional energy markets. IFAC-PapersOnLine. 2020;53(2):12175–82.

31. Wu P yu, Mjörnell K, Mangold M, Sandels C, Johansson T. A Data-Driven Approach to Assess the Risk of Encountering Hazardous Materials in the Building Stock Based on Environmental Inventories. Sustainability (Switzerland). 2021;13(7836):1–26.

32. Von Platten J, Sandels C, Jörgensson K, Karlsson V, Mangold M, Mjörnell K. Using machine learning to enrich building databases-methods for tailored energy retrofits. Energies. 2020;13(10).

33. Wang E, Shen Z, Grosskopf K. Benchmarking energy performance of building envelopes through a selective residual-clustering approach using high dimensional dataset. Energy and Buildings. 2014;75:10–22.

34. Miller C. What's in the box?! Towards explainable machine learning applied to non-residential building smart meter classification. Energy and Buildings. 2019;199:523–36.

35. Yang L, Nagy Z, Goffin P, Schlueter A. Reinforcement learning for optimal control of low exergy buildings. Applied Energy. 2015;156:577–86.

36. Gao X, Malkawi A. A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm. Energy and Buildings. 2014;84:607–16.

37. Su T, Li H, An Y. A BIM and machine learning integration framework for automated property valuation. Journal of Building Engineering. 2021;44:102636.

38. Chen Y, Norford LK, Samuelson HW, Malkawi A. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. Energy and Buildings. 2018;169:195–205.

39. Kropat G, Bochud F, Jaboyedoff M, Laedermann JP, Murith C, Palacios M, et al. Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units. Journal of Environmental Radioactivity. 2015;147:51–62.

40. Seo J, Choi A, Sung M. Recommendation of indoor luminous environment for occupants using big data analysis based on machine learning. Building and Environment. 2021;198(March):107835.

41. Emami Javanmard M, Ghaderi SF, Hoseinzadeh M. Data mining with 12 machine learning algorithms for predict costs and carbon dioxide emission in integrated energy-water optimization model in buildings. Energy Conversion and Management. 2021;238(April):114153.

42. Luo XJ, Oyedele LO, Ajayi AO, Akinade OO. Comparative study of machine learning-based multi-objective prediction framework for multiple building energy loads. Sustainable Cities and Society. 2020;61(May):102283.

43. Towards the development of residential smart districts.pdf.

44. Rehman H ur, Korvola T, Abdurafikov R, Laakko T, Hasan A, Reda F. Data analysis of a monitored building using machine learning and optimization of integrated photovoltaic panel, battery and electric vehicles in a Central European climatic condition. Energy Conversion and Management. 2020;221(March):113206.

45. Pallonetto F, De Rosa M, Milano F, Finn DP. Demand response algorithms for smart-grid ready residential buildings using machine learning models. Applied Energy. 2019;239(February):1265–82.

46. Stoimenov L, Djordjević-Kajan S. An architecture for interoperable GIS use in a local community environment. Computers & Geosciences. 2005;31(2):211–20.

47. Cuca B, Brumana R, Oreni D, Iannaccone G, Sesana M. Geo-portal as a planning instrument: supporting decision making and fostering market potential of Energy efficiency in buildings. Open Geosciences. 2014;6(1):121–30.

48. Vassiliadis P, Karagiannis A, Tziovara V, Simitsis A, Hellas I. Towards a benchmark for etl workflows. 2007;

49. March ST, Hevner AR. Integrated decision support systems: A data warehousing perspective. Decision support systems. 2007;43(3):1031–43.

50. Ahmed A, Ploennigs J, Menzel K, Cahill B. Multi-dimensional building performance data management for continuous commissioning. Advanced Engineering Informatics. 2010;24(4):466–75.

51. Karabegovic A, Ponjavic M. Geoportal as decision support system with spatial data warehouse. In IEEE; 2012. p. 915–8.

52. RALPH KIMBALL JC. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Canada: Wiley Publishing, Inc. 2004;

53. Bédard Y, Merrett T, Han J. Fundamentals of spatial data warehousing for geographic knowledge discovery. Geographic data mining and knowledge discovery. 2001;2(2):53–73.

54. Simitsis A, Vassiliadis P, Dayal U, Karagiannis A, Tziovara V. Benchmarking ETL workflows. In Springer; 2009. p. 199–220.

55. Myers BA. Taxonomies of visual programming and program visualization. Journal of Visual Languages & Computing. 1990;1(1):97–123.

56. Celani G, Vaz CEV. CAD scripting and visual programming languages for implementing computational design concepts: A comparison from a pedagogical point of view. International Journal of Architectural Computing. 2012;10(1):121–37.

57. Johansson T, Vesterlund M, Olofsson T, Dahl J. Energy performance certificates and 3-dimensional city models as a means to reach national targets–A case study of the city of Kiruna. Energy Conversion and Management. 2016;116:42–57.

58. Johansson T, Olofsson T, Mangold M. Development of an energy atlas for renovation of the multifamily building stock in Sweden. Applied Energy. 2017 Oct 1;203:723–36.

59. Eriksson P, Johansson T. Towards Differentiated Energy Renovation Strategies for Heritage-Designated Multifamily Building Stocks. Heritage. 2021;4(4):4318–34.

60. He L, Wu G, Dai D, Chen L, Chen G. Data conversion between CAD and GIS in land planning. In IEEE; 2011. p. 1–4.

61. Macay Moreia J. From DSM to 3D building models: a quantitative evaluation. 2013;

62. Mortensen LF, Tange I, Stenmarck Å, Fråne A, Nielsen T, Boberg N, et al. Plastics, the circular economy and Europe′ s environment-A priority for action. 2021;

63. Baitz M, Kreißig J, Byrne E, Makishi C, Kupfer T, Frees N, et al. Life Cycle Assessment of PVC and of principal competing materials. Commissioned by the European Commission(July 2004). 2004;24.

64. Almasi AM, Zhang Y. Separate collection and recycling of PVC flooring installation residue in Sweden-A system assessment. 2019;

65. Håkansson H, Lindvall J. Livscykelanalys på våtrum: Riktvärden för miljömässig belastning. 2020;

66. Wu PY. Predicting hazardous materials in the Swedish building stock using data mining [Licentiate]. LTH; 2022.

67. Byggindustrier S. Resurs-och avfallsriktlinjer vid byggande och rivning. Kretsloppsrådet; 2017.

68. Neitzel RL, Sayler SK, Demond AH, d'Arcy H, Garabrant DH, Franzblau A. Measurement of asbestos emissions associated with demolition of abandoned residential dwellings. Science of the Total Environment. 2020;722:137891.

69. Franzblau A, Demond AH, Sayler SK, D'Arcy H, Neitzel RL. Asbestos-containing materials in abandoned residential dwellings in Detroit. Science of The Total Environment. 2020;714:136580.

70. Bergmans J, Dierckx P, Broos K. Semi-selective demolition: Current demolition practices in Flanders. In 2017.

71. Lewis M. Incompatible trends-hazardous chemical usage in building products poses challenges for functional circular construction. In IOP Publishing; 2019. p. 012046.

72. Rašković M, Ragossnig AM, Kondracki K, Ragossnig-Angst M. Clean construction and demolition waste material cycles through optimised pre-demolition waste audit documentation: A review on building material assessment tools. Waste Management & Research. 2020;38(9):923–41.

73. Economy C. The circularity gap report—closing the circularity gap in a 9% world. 2019.

74. European Environment Agency. Construction and demolition waste: challenges and opportunities in a circular economy [Internet]. 2020. Available from: doi: 10.2800/07321

75. ECORYS. EU Construction & Demolition Waste Management Protocol. 2016.

76. European Commission. Guidelines for the waste audits before demolition and renovation works of buildings. 2018.

77. Wu PY, Mjörnell K, Mangold M, Sandels C, Johansson T. Tracing Hazardous Materials in Registered Records: A Case Study of Demolished and Renovated Buildings in Gothenburg. In IOP Publishing; 2021. p. 012234.

78. Wilk E, Krówczyńska M, Zagajewski B. Modelling the Spatial Distribution of Asbestos—Cement Products in Poland with the Use of the Random Forest Algorithm. Sustainability. 2019;11(16):4355.

79. Wu PY, Sandels C, Mjörnell K, Mangold M, Johansson T. Predicting the presence of hazardous materials in buildings using machine learning. Building and Environment. 2022;108894.

80. Kling R. Lönsam energieffektivisering : saga eller verklighet? : för hus byggda 1950-75. VVS-företagen; 2012.

81. Brown NWO, Malmqvist T, Bai W, Molinari M. Sustainability assessment of renovation packages for increased energy efficiency for multi-family buildings in Sweden. Building and Environment. 2013 Mar 1;61:140–8.

82. Boverket, Swedish Energy Agency. Basis for the third national strategy of energy retrofitting (Underlag till den tredje nationella strategin för energieffektiviserande renovering). 2019. Report No.: 2019:26, 2019:13.

83. Missing Values in Data [Internet]. Statistics Solutions. [cited 2021 Jul 1]. Available from: https://www.statisticssolutions.com/dissertation-resources/missing-values-in-data/

84. Graham JW. Missing Data Analysis: Making It Work in the Real World. 2008;

85. Baweja C. How to Deal with Missing Data in Python [Internet]. Towards Data Science. [cited 2021 Jul 1]. Available from: https://towardsdatascience.com/how-to-deal-with-missing-data-in-python-1f74a9112d93

86. Amidon A. How to Apply Hierarchical Clustering to Time Series [Internet]. Towards Data Science. [cited 2021 Jul 1]. Available from: https://towardsdatascience.com/how-to-apply-hierarchical-clustering-to-time-series-a5fe2a7d8447

87. Amidon A. How to Apply K-means Clustering to Time Series Data [Internet]. Towards Data Science. [cited 2021 Jul 1]. Available from: https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3

88. Wu PY, Mjörnell K, Sandels C, Mangold M. Machine Learning in Hazardous Building Material Management: Research Status and Applications. Recent Progress in Materials. 2021;03(02):1–1.

BuiltHub

builthub.eu