UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Data-driven recommendations for building energy retrofitting at urban scale

Author:
Oleksandr SHEVCHENKO

Supervisor:
Oleksii PASICHNYI

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2024

Declaration of Authorship

I, Oleksandr SHEVCHENKO, declare that this thesis titled, "Data-driven recommendations for building energy retrofitting at urban scale" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Data-driven recommendations for building energy retrofitting at urban scale

by Oleksandr SHEVCHENKO

Abstract

Decreasing the number of retrofitting recommendations based on building stock data without using expensive and computational-consuming UBEM models or energy advisors' work could allow upscale of retrofitting decision-making for cities or districts and also for smaller units, such as associations of property owners or development companies.

Despite the extensive amount of research in building rennovaiton area, the question of decreasing the number of retrofitting measures, that should be validated for every single building is commonly out of the research interests area.

This work aims to investigate an approach to identify feasible retrofitting recommendations for existing building stock using data-driven approaches and machine-learning techniques based on urban-level datasets.

We approached the problem as a classification task using the Swedish EPCs dataset as data input and created a dataset for using retrofitting measures recommendation from EPCs declarations as classification labels.

In this study, we tested multi- and single-label classification approaches and various classification algorithms. Results confirmed that this research area is promising, but obtained classification performance is insufficient for the industry usage. The binary classification on single retrofitting measures achieving high precision, but low recall. This makes this method possible to be used in the task of enhancing building stock datasets with missing retrofitting measures.

Acknowledgements

I am deeply thankful to my supervisor, Oleksii Pasichnyi, from KTH Royal Institute of Technology. His guidance, support, and valuable insights have been instrumental in this master's thesis.

I am also grateful to AresAI for generously providing computational resources, which significantly aided in the execution of this investigation.

I want to thank Ukrainian Catholic University, Oleksii Molchanovskyi, Ruslan Partsey, and the Faculty of Applied Sciences for the exceptional Master's Program in Data Science. Their guidance and support have not only expanded my knowledge but also opened doors to the world of Data Science.

Lastly, I want to extend my gratitude to my wife and my daughter. Their endless support has been my rock throughout this whole path. I couldn't have done it without their patience and understanding. They have sacrificed so much precious time we could have spent together. Thank you!

Contents

Declaration of Authorship ii

Abstract iii

Acknowledgements iv

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Research Objective . . . 2
- 1.3 Structure of the thesis . . . 3

2 Literature review 4
- 2.1 Data-Driven and Machine Learning Approaches in Building Retrofitting Recommendations . . . 4
- 2.2 Data sources in energy and building retrofitting modeling . . . 7
  - 2.2.1 Conclusion . . . 8

3 Methodology 9
- 3.1 Research Gap and Problem Formulation . . . 9
- 3.2 Research Setting and Approach to Solution . . . 10
- 3.3 Evaluation . . . 11
- 3.4 Conclusion . . . 12

4 Dataset 13
- 4.1 EPC data cleaning . . . 13
  - 4.1.1 Overview of Energy declarations' data . . . 13
  - 4.1.2 Forming research dataset . . . 15
- 4.2 Classification datasets . . . 16
- 4.3 Conclusion . . . 17

5 Experiments 19
- 5.1 Multilabel classification . . . 19
  - 5.1.1 Classification on the full dataset . . . 21
  - 5.1.2 Classification on the split data . . . 22
    - Spliting by year of construction . . . 22
    - Splitting by building category . . . 24
  - 5.1.3 Conclusion . . . 25
- 5.2 Binary classification . . . 26
- 5.3 Experiments limitations discussion . . . 28

6 Conclusions 30
- 6.1 Discussion . . . 30
- 6.2 Future work . . . 31

A Appendix: dataset structure **32**
A.1 Structure of Swedish EPCs dataset 32
A.2 Statistics of classification dataset 42

B Appendix: Results of experiments **45**
B.1 Multilabel classification results on split data 45
B.1.1 Classification results on data split by year of construction . . . 45
B.1.2 Classification results on data split by building category 47
B.2 Single-label classification results 48

Bibliography **54**

List of Figures

1.1 Selecting feasible retrofitting measures for the specific building (Ma et al., 2012) 2

4.1 Number of Energy declarations by Year 14
4.2 Distribution of Number of Buildings with Retrofitting Recommendations to Buildings without Recommendations by retrofitting measures 17
4.3 Number of building retrofitting recommendations in comparison with the implemented retrofitting measures 17

List of Tables

2.1 Overview of the data sources and approaches used for retrofitting recommendations . . . 5

4.1 Energy declarations dataset general structure . . . 14

5.1 Multilabel classification: result summary . . . 21
5.2 Multilabel classification: Classification results on data split by year of construction. F1-score values for different methods and datasets. DS 1 stands for Dataset 1, and DS 2 stands for Dataset 2 . . . 23
5.3 Multilabel classification: Classification results on data split by building category. F1-score values for different methods and datasets. DS 1 stands for Dataset 1, and DS 2 stands for Dataset 2 . . . 24
5.4 Binary classification: results summary, best classifier and dataset for each retrofitting measure . . . 27
5.5 Single-label classification: aggregated classification performance score . . . 27

A.1 Retrofitting measures recommendations columns in a dataset with artificial recommendation code and recommendation name in English . 32
A.2 Data categories encoding . . . 33
A.3 Energy declarations dataset structure . . . 33
A.4 Statistics of classification dataset . . . 42

B.1 Multilabel classification: results' summary for building built before 1920 45
B.2 Multilabel classification: results' summary for building built between 1920 and 2000 . . . 46
B.3 Multilabel classification: results' summary for building built after 2000 46
B.4 Multilabel classification: results' summary for building category "Multi-family dwellings" . . . 47
B.5 Multilabel classification: results' summary for building category "Office buildings" . . . 47
B.6 Multilabel classification: results' summary for building category "Single- or two-family house" . . . 48
B.7 Single-label classification: results' summary for LogisticRegression Classifier . . . 48
B.8 Single-label classification: results' summary for KNeighbors Classifier 49
B.9 Single-label classification: results' summary for DecisionTree Classifier . . . 50
B.10 Single-label classification: results' summary for RandomForest Classifier . . . 51
B.11 Single-label classification: results' summary for ExtraTree Classifier . . 51
B.12 Single-label classification: results' summary for ExtraTrees Classifier . 52

List of Abbreviations

ANN	**A**rtificial **N**eural **N**etwork
BEM	**B**uilding **E**nergy **M**odelling
BES	**B**uilding **E**nergy **S**imulation
EPC	**E**nergy **P**erformance **C**ertificates
MFD	**M**ulti-**F**amily **D**wellings
OB	**O**ffice **B**uildings
SFH	**S**ingle- or two-**F**amily **House**
UBEM	**U**rban **B**uilding **E**nergy **M**odelling

Chapter 1

Introduction

1.1 Motivation

The United Nations' Sustainable Development Goals emphasize the critical role of improving building energy performance in global energy and emissions reduction efforts. Urban environments are the largest energy consumers globally (Deb and Schlueter, 2021), and use up to 70% of all primary energy (Johari, Shadram, and Widén, 2023).

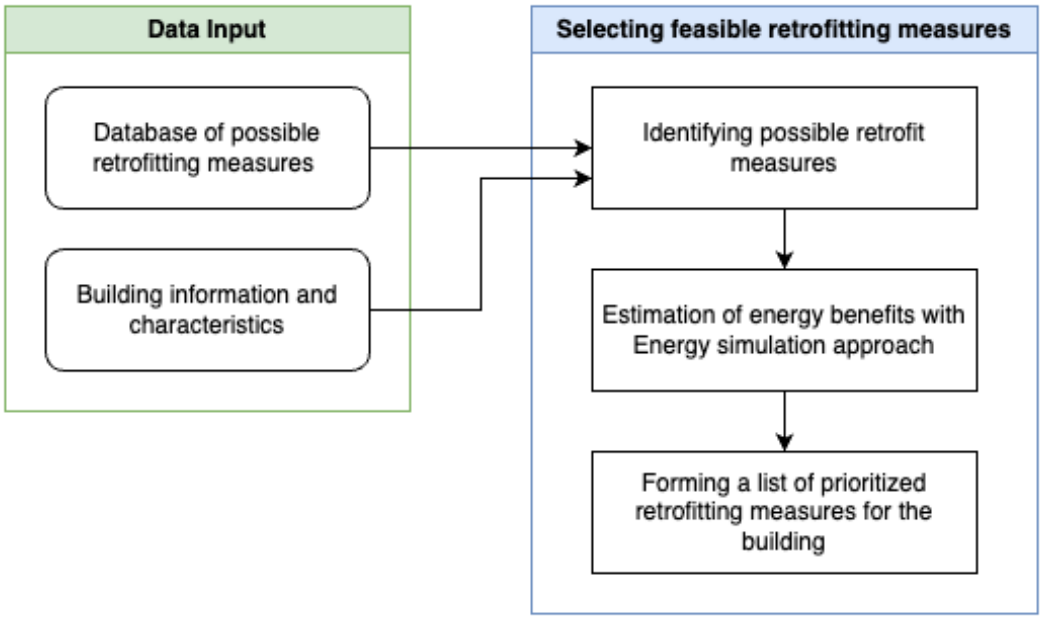
To reach energy and climate goals on the city level, there is a pressing need to enhance the buildings' energy efficiency and reduce related greenhouse gas emissions. In Europe, the aging building stock, with over 35% of buildings being over 50 years old (Ali et al., 2020), what further complicates this as only a tiny percentage undergo retrofitting annually and a significant part of the current global building stock will still be in use by 2050 (Deb and Schlueter, 2021). The steady increase in energy use in buildings, especially in developing countries, highlights the urgency of more efficient building retrofit methodologies. This approach is a key to meeting the Paris Agreement's (The Paris Agreement, 2015) emission-reduction targets.

Upgrading existing dwellings presents several challenges, including the high costs involved, the complexity of determining the most effective intervention strategies, and the execution difficulties associated with large-scale retrofit projects. Development of a renovation plan for the specific building may be represented as a process that includes choosing relevant energy conservation measurements from all possible options that can be applied to the building, estimating the impact from an energy-saving standpoint, and validating each option as optimal or not based on the impact and renovation constraints, such as costs, feasibility for the specific building characteristics, etc (Ma et al., 2012). The process of selecting a feasible list of retrofitting measures for the specific building, described by Ma et al., 2012, is schematized in Figure 1.1

The energy simulation approach is commonly used to choose a feasible list of retrofitting measures from all possible ones for the building. Building Energy Modeling (BEM) for individual buildings is a way to simulate energy consumption based on building characteristics. Urban Building Energy Modelling (UBEM) tools allow energy demand simulation on a large scale (Ferrando et al., 2020). Such methods focus on energy modeling and consider retrofitting recommendations that can be estimated within a specific modeling tool and validate chosen recommendations using computational approaches. This may lead to over-generic recommendations, especially on the urban scale, due to the high uncertainty factors in the modeling process.

The scaling question is important not only for the urban level but also for the availability of the recommendation process for individual building owners. Due to the complexity and costs of creating an individual building model and especially

FIGURE 1.1: Selecting feasible retrofitting measures for the specific building (Ma et al., 2012)



upscaling it to the urban level, testing retrofitting recommendations utilizing BEM or UBEM processes is far from optimal.

This leads to the conclusion that the step "Identifying possible retrofit measures" (Figure 1.1) is a crucial part of providing applicable retrofitting recommendations, specifically on a large scale. Validation of all possible retrofitting measures is not effective since some of them cannot be applicable to the specific building. A common practice to form the list of possible retrofit measures is either using the measures supported by the selected energy emulation tool or forming the list of possible retrofit measures can be done by energy advisors. As an expert work, the results provided by energy advisors are more accurate, but the throughput of the advisors is insufficient to handle all the building stock. According to the "In-depth follow-up municipal energy and climate advice, 2022"[1], not more than 25% of all incoming recommendations were handled.

1.2 Research Objective

This thesis aims to optimize the step "Identifying possible retrofit measures" (Figure 1.1) and develop an approach to identify possible retrofitting recommendations based on the building information and characteristics at the urban scale. Decreasing the number of possible retrofitting options will optimize the process of choosing feasible measures both for the energy modeling approach or by energy advisors - fewer input retrofitting options for validation means increasing the number of handled declarations.

Taking Stockholm building stock as a case study, we propose a data-driven approach that takes the same UBEM data as input and a classification approach to choose possible retrofitting recommendations based on the available list of recommendations from Swedish Energy Performance Certificates (EPC) data.

[1] *In-depth follow-up municipal energy and climate advice (EKR)* 2022.

We formulated the set of research questions that we will investigate in the experimental part of our work:

- Validate that the classification approach is applicable to choosing possible retrofitting measures based on the building stock data without using the building energy demand evaluation
- Identifying the whole set of retrofitting recommendations for the specific building using multi-label classification methods
- Identifying a single retrofitting recommendation using classification methods
- Validate the hypothesis that splitting data based on natural criterion may improve the classification performance

1.3 Structure of the thesis

In Chapter 2, we review related works in the area of building stock retrofitting. Chapter 3 includes a gap analysis, problem definition, and proposed solution alongside experiment and evaluation setting. Chapter 4 contains the initial data overview and the process of forming datasets for experiments with their main characteristics. Chapter 5 outlines a detailed overview of the experiments. Chapter 6 concludes the work by summarising our results and outlining possible future research directions.

Chapter 2

Literature review

Developing renovation strategies for existing building stock is a multifaceted task that can be considered from different research areas' perspectives. In the scope of this research, we focused on data-driven approaches and machine-learning techniques used for retrofitting planning. Since building energy modeling (BEM) and Urban Building Energy Modeling (UBEM) commonly focus on building retrofitting tasks, specifically an impact estimation, research studies in these areas cannot be separated from building retrofitting recommendations development completely. At the same time, this research implies the usage of data input similar to the one utilized in UBEM, specifically EPC data. We will omit details of UBEM approaches, and while analyzing building energy modeling studies will primarily focus on the UBEM input data.

2.1 Data-Driven and Machine Learning Approaches in Building Retrofitting Recommendations

Retrofitting recommendations analysis is a similar process to building energy modeling. Grillone et al. (2020), in their review distinguish three main groups of methods to predict energy demand or retrofitting strategies planning: Deterministic methods, which incorporate building energy simulation (BES) approaches with modifications; Hybrid methods represented BES in combination with data-driven techniques and Data-driven approaches focused on providing recommendations based on collected data. Approaches used in reviewed studies focus on providing retrofitting recommendations using energy measurements as criteria to identify retrofitting recommendations.

With a widespread of machine learning techniques, they are often used in energy modeling and retrofitting planning. Shu and Zhao (2023), in their review, listed the most common approaches for decision-making in urban retrofitting for energy performance optimization. According to this review, Energy simulator-based, and Optimization modeling-based approaches are the most common. They are based on the usage of Energy simulation software or custom simulation or optimization models, including analytical techniques and machine learning approaches. In the review of data-driven approaches in building energy retrofitting, Alrobaie and Krarti (2022) mentioned that the popularity of data-driven techniques has steadily grown over the last decade. They analyzed the following data-driven approaches and ML techniques and their applications for building energy consumption predictions: Linear Regression, Decision Tree, and Ensemble Methods, including Random Forest, Gradient Boosting Machine, Support Vector Machine, Artificial Neural Network, and Kernel Regression.

Despite the fact that Building Energy Modeling (BEM) techniques are commonly used to define suitable retrofitting solutions, there are shortfalls in them, which provide room for improvements. Pan et al. (2023) investigated approaches and methods used to evaluate building energy performance on individual and urban levels. The general drawbacks of approaches based on simulation techniques are the requirement of detailed buildings' physical characteristics, a large amount of energy data and related data, and complexity in model development.

Analyzing studies in building retrofitting area we can conclude that machine-learning techniques have become more and more common. However, identifying retrofitting plans using energy modeling is a common approach. Thrampoulidis et al. (2021) proposed a surrogate Artificial neural network-based model to predict retrofit solutions for residential buildings, which decreased computational time ten times. In further research, this model was used on a large scale to provide optimal retrofit solutions for Switzerland's building stock (Thrampoulidis, Hug, and Orehounig, 2023). Biessmann, Kamble, and Streblow (2023), in their study, used the AutoML model to predict energy consumption and energy saving of large public buildings. Classic linear approaches are applicable as well. For instance, Pedone et al. (2023) proposed a framework based on a multi-linear prediction model to plan energy refurbishment of school buildings in Italy.

Ensembling different ML models is also regularly used. Zhang et al. (2022) proposed a data-driven framework to find an optimal retrofitting plan for a specific building. They developed an Artificial neural network (ANN) model to predict building energy performance with different retrofitting packages and used a multi-criteria decision-making algorithm to choose optimal retrofitting recommendations. To mitigate the lack of data or low data quality, different techniques of datasets enhancement are used. For instance, Feng et al. (2022) in their study proposed an approach for retrofit strategy analysis based on the data imputation method and energy performance calculation with Bayesian neural network and Fuzzy C-means clustering for Swedish building stock.

TABLE 2.1: Overview of the data sources and approaches used for retrofitting recommendations

Input data	**Method for retrofitting measures' selection**	**Approach**	**Features validation**	**Reference**
Building characteristics	Supported by used UBEM models	Surrogate ANN-based ML model	Only model results validation	Thrampoulidis et al., 2021
Building archetypes Climate data	Supported by building simulation tool	ANN models ensemble	Only model results validation	Thrampoulidis, Hug, and Orehounig, 2023
Building features Yearly aggregated climate data	Used energy consumption optimization coefficient without specific features	AutoML model	Energy modeling validation only	Biessmann, Kamble, and Streblow, 2023

Building characteristics	Selected from possible recommendations according to New York City's Local Law 87	FRL classification model	ROC AUC score	Marasco and Kontokosta, 2016
Building characteristics	Supported by building energy modelling tools HOT2000 and HTAP	ANN model for energy consumption evaluation Multi-criteria decision-making algorithm	Pareto optimal retrofit solutions by reducing carbon emission and lyfe cycle costs	Zhang et al., 2022
Geometry data Building thermal properties Weather data	Predefined features based on earlier researches	Multi-linear regression model	Prioritise retrofitting scenario based on energy performance tool simulation results	Pedone et al., 2023
EPC data	Retrofitting strategies from Swedish EPC	Bayesian regularization backpropagation neural networks (BRBNNs) Fuzzy C-means clustering (FCM) for performance modelling	Energy modeling results	Feng et al., 2022
Building properties	Retrofit recommendations for EPC rating improvement	Gradient boosted regression tree	Estimate building energy performance	Seyedzadeha et al., 2020
Building stock data, EPC Weather data Census	Based on EPC data	Ensemble Gradient Boosting algorithm (XGB, LGBM, and HGB)	Buildings energy modeling apporach	Ali et al., 2024

2.2 Data sources in energy and building retrofitting modeling

Choosing the proper data is a crucial and fundamental step. In building retrofitting modeling, a common approach is to mix data sources and create extended datasets for modeling. For instance, Ali et al. (2020) utilized a complex approach in proposed data-driven retrofit modeling for Dublin's building stock. Their methodology contains various input data types, including geometric and non-geometric factors, building energy performance indices, statistical data, retrofit measure costs, and retrofit expert reports.

Common examples of possible data sources includes, but not limited to

Building stock databases and analysis projects. A range of European and international level projects are dedicated to the collection and analysis of building stock data. Ali et al. (2019) in their review listed such initiatives like *ODYSSEE-MURE* (2019), *Eurostat* (2019), *TABULA* (2019), *ENTRANZE* (2019), and *BPIE* (2019). For instance, *TABULA* (2019) - The Typology Approach for Building Stock Energy Assessment - aimed to create a comprehensive database encompassing various European building typologies.

Surveys, census, and metering data. According to Ali et al. (2019), census and survey data are crucial resources for data collection within building stock analysis. Metering data is also often used for building stock clusterization.

Climate data Climate data represents the location-specific data. This kind of input includes air temperature, relative humidity, wind, diffuse and direct solar radiation, and other similar information that can be obtained from meteorological databases and geographical information systems.

Energy Performance Certificates Improving building energy performance is a key focus of the EU's energy and climate policy. The Energy Performance of Buildings Directive (EPBD), initiated in 2002[1] and revised in 2010[2], is the primary mechanism for this improvement. The directive covered various strategies, including new building codes, energy retrofitting of existing building stock, creating financial incentives for energy efficiency, and influencing consumer behavior. A central component of the EPBD is the Energy Performance Certificate (EPC). In the EPBD 2022 the EU Parliament defined EPC as "a certificate recognized by a Member State or by a legal person designated by it, which indicates the energy performance of a building or building unit". EPC contains detailed information on a building's energy use, including data on building reference, geometry, audit methodology, energy consumption, system installations, and energy efficiency recommendations.

Dahlström, Broström, and Widén (2022) mentioned EPC as an essential resource in the EU, which is used for real-estate analysis projects, national and city-level decision planning, and the buildings energy modeling areas. According to Pasichnyi et al. (2019) EPC data is commonly used in most building energy modeling related research studies and projects. In the review of 79 papers thirteen possible problem domains were identified, where EPC databases can be applied, including evaluation

[1] *DIRECTIVE 2002/91/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL* 2002.
[2] *DIRECTIVE 2010/31/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL* 2010.

and predictions of energy demand, urban planning, building design, retrofitting, etc. For instance, Ferrantelli and Kurnitski (2022) tested different methods for energy performance labeling using data from the Estonian EPC database. Applying different EPC labeling methods prescribed by the EU, they estimated retrofit and renovation rates, CO2 emissions reduction, and energy saving.

2.2.1 Conclusion

In this section, we presented a review of used data-driven and machine-learning approaches as well as used input data overview for providing retrofitting recommendations and evaluation of building energy demand. Table 2.1 provides an overview of data inputs and retrofitting recommendations approaches in reviewed studies. Most of the studies do not focus on the process of selecting possible retrofitting measures before energy modeling and using energy modeling approach to form a list of prioritised retrofitting options.

Chapter 3

Methodology

3.1 Research Gap and Problem Formulation

In current practice, the decision-making about building retrofitting is quite a challenging task that contains several stages, including selecting applicable and feasible retrofit options based on the specific building characteristics, modeling the effect of chosen retrofitting options, and figuring out the optimal renovation actions solutions according to goals and limitations. Building stock retrofitting is affected by many uncertainty factors, like building-specific information, retrofit technologies, human-behavior factors, climate, and many others. A good estimate of uncertainty factors is crucial for choosing the most effective retrofitting options aimed at maximizing a building's energy efficiency throughout its entire lifespan (Ma et al., 2012).

Most of the reviewed studies use the modeling of building energy demand to evaluate the quality of provided recommendations (Table 2.1). This requires developing a detailed physics building model or UBEM usage for large scales like cities or districts (Ali et al., 2024). These approaches are computationally consuming, and creating such models is meticulous work. Additionally, validation of the selected retrofitting features is often done by the energy demand simulation result. The step of selecting the feasible retrofitting options for the specific building or building stock is commonly skipped or does not include the analysis of the building stock but is based on the computational and simulation model characteristics and possibilities. This leads to two possible edge cases - when retrofitting options might be either not considered but is feasible or modeled and estimated but not feasible for the specific building.

Decreasing the amount of applicable retrofitting options is commonly skipped but is a viable step in the building renovation strategies development process. For a specific building, this can reduce computational time and resources for energy simulation tools. Energy modeling on a large scale often uses the concept of building archetypes or reference buildings, which represent a group of similar buildings stock. In this case, choosing a feasible set of retrofitting options becomes essential, especially when not all retrofitting scenarios can be implemented for every real building.

Defining the possible list of recommendations is quite challenging and often requires an energy or building audit (Ma et al., 2012). One of the possible solutions is using retrofitting scenarios provided by national or urban levels. For instance, Marasco and Kontokosta (2016) used energy conservation measurements based on US Energy databases, Ali et al. (2024) based the selection of discovered retrofitting options on Ireland's EPC data.

Reducing the number of possible retrofitting approaches before building audit or energy modeling will reduce the cost and time spent on development renovation strategies. To solve this, we aim to develop a data-driven approach that will

be able to identify possible retrofitting measures based on the building information and characteristics at the urban scale. This short list of retrofitting measures will be further used to form a list of prioritized retrofitting recommendations by energy modeling. Figure 1.1 provides the general process schema.

Considering the research questions outlined in Section 1.2, we set the following goals:

- for prediction of the whole set of retrofitting measures for the building - correctly identify more than 90% of feasible retrofitting measures
- for prediction of the single retrofitting measure - identify the retrofitting measure for more than 90% of the tested building stock

3.2 Research Setting and Approach to Solution

We outlined the aim of our research as identifying possible retrofitting measures for the specific building from the whole set of retrofitting measures based on the building information and characteristics. The important point here is that there is no aim to provide a final recommendations list for the building but to narrow down the list of all possible retrofitting measures that can be estimated for the building to the list of feasible for this building. All further usage of retrofitting recommendations term means the applicable list of retrofitting measures for the building. Further evaluation of the feasible retrofitting measures and providing a prioritized list of retrofitting recommendations are not in the scope of this research.

Approach to Solution We frame our primary task as using a binary classification approach to identify retrofitting recommendations for buildings at an urban scale based on the building stock data. We take Sweden Energy Performance Certificates (EPC) data as source data about the building stock. EPC data in Sweden contain over thirty possible recommendations for retrofitting measures to improve building energy performance, Section X in Table A.3. They are divided into three main groups: Building energy efficiency enhancements, infrastructure upgrades, and control and efficiency technologies. These recommendations are represented as binary values, so the value for each retrofitting measure recommendation can be either True or False. The binary nature of the recommendations values allows us to consider these recommendations as labels assigned to buildings and apply classification methods to building stock data to predict these labels.

We start with cleaning the raw EPC data and forming two dataset versions for further processing. We then experiment with multi- and single-label classification methods. Also, we tested the natural-based data split approach to improve performance.

We conclude experiments for predicting retrofitting recommendations by comparing the results obtained in the same setting. We conclude our work by discussing the results and the directions for further research.

Experiment Setting

Compiling datasets After initial data cleaning and forming a base dataset from the raw EPC data, we compile two dataset versions with different representations of selected features. EPC data contains information about both implemented retrofitting

measures and retrofitting recommendations. We will use retrofitting recommendations as binary labels (Section X in Table A.3), and information about implemented retrofitting measures (Section IX in Table A.3) will be included in building info.

Data spliting Building stock data contains several splitting parameters that allow data to be distinguished into groups and to test the research hypothesis that natural data splitting may improve classification performance.

There are two splitting criteria:

- by the building category - based on the idea that constructionally different types of buildings require different retrofitting options, for instance, small one- or two-family houses versus dwellings higher than two floors
- by the year of construction - this split is driven by the historical nuances of the development of the housing market in Sweden and changes in building technology in 19s, 20s, and 21st centuries

We will apply classification algorithms to the split data and check classification performance compared to classification performance on the full dataset and between two splitting groups.

Multilabel classification We will start by considering the task a binary multilabel classification task. We will use logistic regression as a baseline classifier. It's simple enough and require minimum hyperparameter tuning. After defining the baseline we will move to K Nearest Neighbors, Decision tree algorithm, Random Forests for Multi-Label. We will test Classifier Chain method alongside with mentioned algorithms to identify how it affects the performance.

Binary classification Next, we will test an approach where each retrofitting recommendation is considered as an independent binary label. We will train a separate classifier model independently for every retrofitting measure. Similar to the multilabel approach, it will use logistic regression as a baseline classifier. Next, we will apply the K Nearest Neighbors algorithm. Then, we will move to tree classifiers - Decision tree, Extra trees, and Random Forests.

3.3 Evaluation

Multilabel Classification Given the complexity and label imbalance in our dataset, we utilize metrics that better capture the performance across all labels.

We use the **F1-Score Average Samples** approach to evaluate the performance of the multilabel classification. This version of the F1-score calculates metrics individually for each instance and then averages them, thus emphasizing the performance on a per-sample basis rather than per class. This is critical in our multilabel setting, where labels are highly imbalanced.

The formula for the samples-averaged F1 score is:

$$F1_{samples} = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} 2 \cdot \frac{p_i \cdot r_i}{p_i + r_i}$$

Where:

- n_{samples} is the total number of samples

- p_i is the precision for the i-th sample
- r_i is the recall for the i-th sample

This method gives more weight to samples with more labels, as they contribute more to the overall performance of the model. (Gibaja and Ventura, 2015; Tsoumakas, Katakis, and Vlahavas, 2010)

We use **Hamming Loss** to quantify the fraction of incorrect label predictions over the total number of labels.

Binary Classification We use the **F1-Score** as the main metric. This metric is highly effective for imbalanced datasets due to its sensitivity to the balance between precision and recall.

We use **Precision and Recall** metrics to collect more information about the model performance. They provide additional information about the prediction quality and are useful in case of an imbalanced dataset.

We do not use **Accuracy** due to the misleading nature of accuracy in imbalanced datasets, where it may disproportionately reflect the majority class's prevalence. This approach aligns with best practices in evaluating classifiers under skewed class distributions.

3.4 Conclusion

Reducing the number of possible retrofitting measures before conducting an audit or energy modeling can significantly decrease the needed time and resources. Our research aims to address this by developing a data-driven approach that identifies feasible retrofitting measures based on building characteristics at an urban scale. In our research, we will use classification methods to identify feasible retrofitting measures and Sweden Energy Performance Certificates data as source data about the building stock. We will compile a research dataset, considering the provided in EPC data retrofitting measures as ground truth.

Chapter 4

Dataset

The data source for this work is the energy declarations' dataset provided by Boverket (Swedish National Board of Housing)[1] and obtained by the supervisor from KTH. This dataset contains information about buildings across Sweden and is allowed to be used for research purposes[2].

The research version had these alterations done by Boverket before export. Declarations with the following criteria are excluded from the withdrawal:

- Buildings that are not subject to declaration

  declarations with an area less than 50 m^2

- Unreasonable values

  declarations with energy performance less than 20 kWh/m^2 per year

  declarations with energy performance greater than 500 kWh/m^2 per year

  declarations with building category One- and two-dwelling houses and an area larger than 500 m^2

The dataset was provided in Swedish, and the translation of the dataset was not considered a part of this work. This leads that only some data entries will be translated into English, but generally, columns and terms may be shown in Swedish.

4.1 EPC data cleaning

4.1.1 Overview of Energy declarations' data

The structure of the original data is present in Table 4.1 and replicates the structure of the building energy questionnaire form. Important features of the original data are:

- Splited into several files by year from 2019 to 2023
- Contains duplicated records, which may be as data adjustments as well as new historical updates of the declaration data
- Single declaration may have several sets of retrofitting recommendations
- Some items may not contain information about retrofitting measures
- Contains several records for the same address, for instance, when several buildings share the same address

[1] https://www.boverket.se/
[2] https://www.boverket.se/sv/energideklaration/energideklaration/

- recommended retrofitting measures were provided by humans (energy advisors) and were not additionally verified so that they may contain mistakes
- retrofitting measures contains not standardized columns "Other" for every group of retrofitting options, which are not detailed and cannot be processed

The full structure of Building Energy declaration data with columns, selected for further processing and classification task described in Table A.3

FIGURE 4.1: Number of Energy declarations by Year

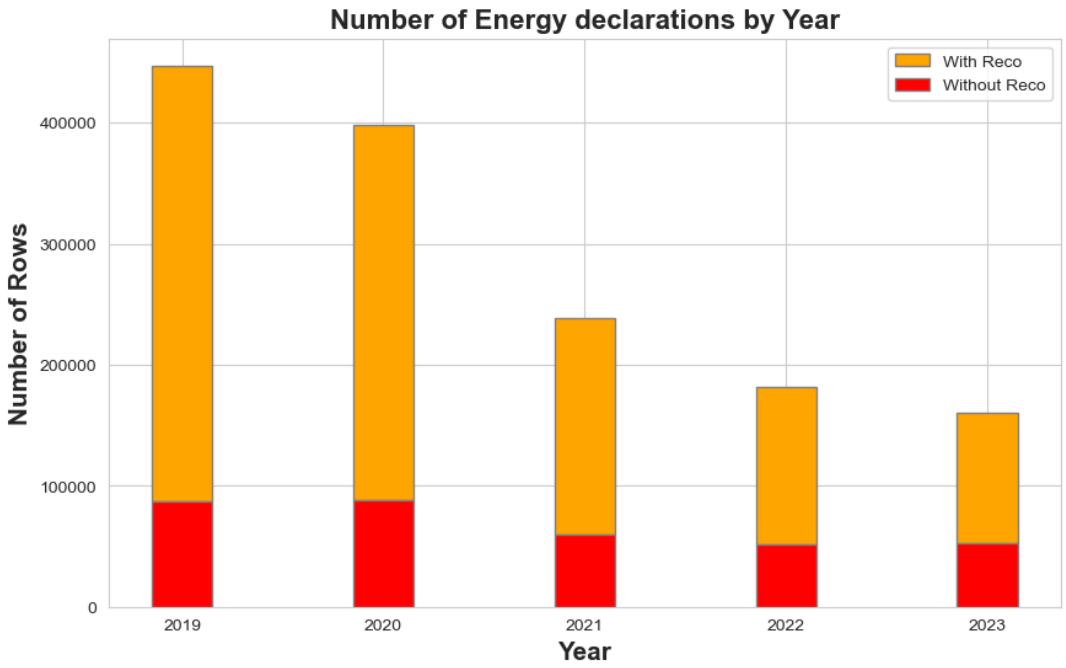


TABLE 4.1: Energy declarations dataset general structure

Section	Data in section	Total number of columns	Number of columns selected for classification
I. The building - Identification	Building information Building address Real estate data Building address	12	0
II. The building - Properties	Energy Declaration data	31	15
III. Energy use	Energy Declaration data	47	31
IV. Information about radon	Energy Declaration data	4	2
V. Information on ventilation control	Energy Declaration data	9	6
VI. Air conditioning system details	Energy Declaration data	1	1

Continuation of Table 4.1

Section	Data in section	Total number of columns	Number of columns selected for classification
VII. Inspection of heating systems	Energy Declaration data	10	1
VIII. Air conditioning system inspection	Energy Declaration data	10	1
IX. Carried out energy efficiency measures since the previous energy declaration	Retrofitting measures Energy Declaration data	35	28
X. Recommendations on cost-effective measures	Retrofitting measures Energy Declaration data	38	28
XI. Miscellaneous	Energy Declaration data	2	0
XII. Expert	Energy Declaration data	2	0
XIII. Building's energy performance	Energy Declaration data	3	0

4.1.2 Forming research dataset

The original data represents exported information from the Building Energy declaration surveys. We start with data-cleaning of the provided data, and this process contains several steps

Handling multiple recomendations Due to the nature of the EPC data export format, the data may contain several sets of retrofitting recommendations for the same building. All of them are applicable to the building, and we merge them into a single set of retrofitting measures.

Processing several buildings within the same address For records that belong to the same address, we choose the one that is marked as the main address. Other records will be marked as duplicated and not included in the dataset.

Dealing with historical updates Another case when data contains several rows for the same building is the declaration updates. It may happen due to fixing errors or adding missed information to the same energy declaration, or providing a new energy declaration for the building. We distinguish these cases using the period between declarations in dataset records. When there's less than one year between sequential declarations we consider the last one as a correction, otherwise - as a new version of declaration.

The total number of identified historical updates is about 2300 declarations, which is less than 1% of the total number of declarations in the researched dataset. They wouldn't provide significant differences but may cause bias due to similar building characteristics. We decided not to include historical updates to the research dataset.

Data cleaning After combining all data we do the cleaning stage

- fill empty values with appropriate values due to the physical meaning

  converting text yes/no fields into binary

  one-hot encoding for categorical data

- fix incorrect values in data

  fix number of floors that are physically nonsense, for instance more than five floor for one- or two-family houses

  limit number of basement floors up to two - this is common Sweden limitation

  set minimum physical limit of airflow

4.2 Classification datasets

In the previous step, we created a dataset that will be a base for the specific classification datasets. Main characteristics of dataset

- 274878 rows and 110 columns
- 28 columns contain retrofitting recommendations - these will be our prediction labels
- 28 columns contain implemented retrofitting measures and will be used as features
- 54 columns contain building characteristics and information about the energy usage

The number of retrofitting recommendations is small enough to conclude that the dataset for each single label and between labels is also imbalanced. Figure 4.2 provides a graphical overview of the distribution of buildings with- and without retrofitting recommendations for each retrofitting measure. For every single measure, the total number of buildings with retrofitting recommendations is not more than 20% compared to the total number of buildings. Measures R14, R17, R21, R24 have the biggest number of buildings with recommendations, for others number of buildings with retrofitting recommendations do not extend 10%.

We encoded retrofitting recommendations to provide a way of showing and mentioning specific retrofitting measures. The full list of recommendation codes with original column names and English-translated recommendation names is available in Table A.1.

Building retrofitting recommendations and the implemented retrofitting measures are also highly imbalanced among themselves.

Another important data characteristic is that all numeric distributions are skewed and contain natural outliers.

More detailed statistics for the numeric columns of the dataset are present in Table A.4. Due to the realistic nature of the outliers in datasets, we decided not to remove them but to mitigate their effect by considering energy usage values with respect to the building area.

For the experiments, we created two versions of the classification datasets. In these versions, columns from the "Energy use" section of the Energy declaration datasets are processed differently.

FIGURE 4.2: Distribution of Number of Buildings with Retrofitting Recommendations to Buildings without Recommendations by retrofitting measures

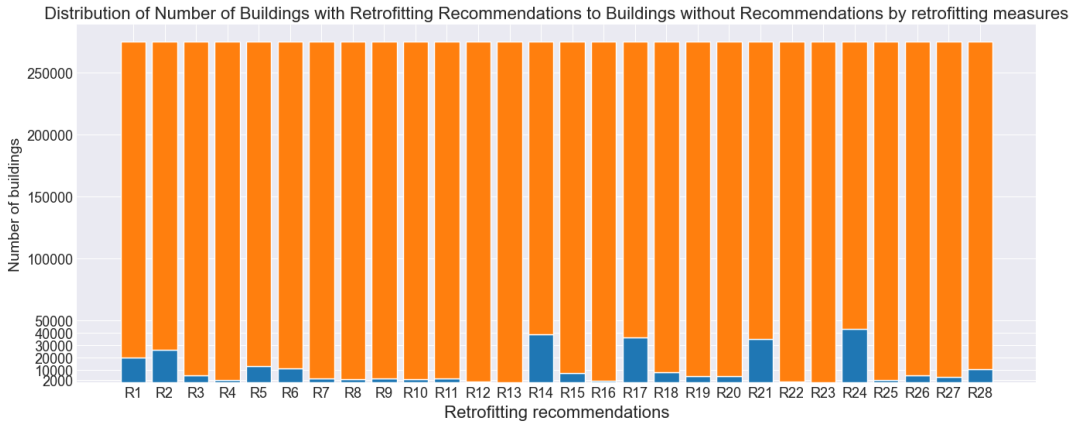
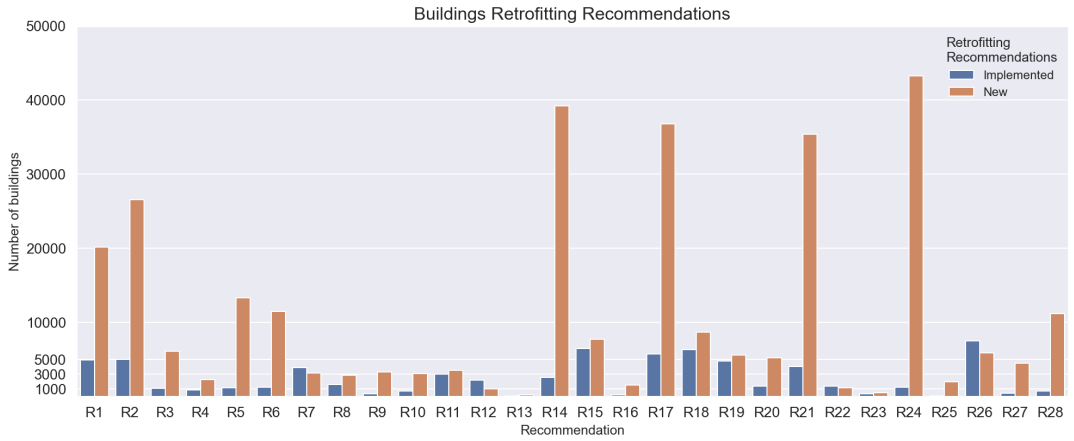


FIGURE 4.3: Number of building retrofitting recommendations in comparison with the implemented retrofitting measures



Dataset 1: detailed energy usage The value of each of the energy usage columns is divided by building area, so we have energy usage per square meter by specific source - heating, hot water, electricity, and their subtypes.

Dataset 2: total energy usage by type In this case, we sum each building's energy usage by type: total usage, heating, hot water, and electricity. These total values are divided by building area, so in classification, we deal with total energy usage by type per square meter. The energy columns are converted to binary.

4.3 Conclusion

In this chapter, we did cleaning of the original Sweden EPC data and compiled the research dataset. We created two versions of the dataset for further experiments. Each of these versions contains 74878 rows and 110 columns; 28 of the columns are retrofitting measures that will be used as prediction labels.

Despite the cleaning and feature selection process, the research dataset still has limitations:

- Retrofitting options were not additionally verified and so that may be inaccurate
- "Other" retrofitting recommendations were excluded, which may lead to missing some important data dependencies or missing correct prediction labels

Chapter 5

Experiments

This chapter outlines our experiments. We start with data analysis and creating two dataset versions for validation of the classification method for providing retrofitting recommendations. Dataset characteristics are discussed in Section 4.2. We start experiments with a multilabel classification approach in Section 5.1 and test different classification algorithms and approaches on two dataset versions. We continue with the validation hypothesis about possible performance improvements on split data, using two approaches: split by year of construction and by building type. We summarise our findings in Section 5.1.3. We conclude our experiments by testing binary classification methods aims to predict every single retrofitting measure separately and discussing the results in Section 5.2. Limitations of the processed experiments discussed in the Section 5.3

5.1 Multilabel classification

At this stage of our experiments, we tested the multilabel classification approach, where each retrofitting measure recommendation was treated as a distinct label within our dataset. These experiments aimed to assess the applicability of multilabel classification with complete datasets and data split into groups to predict the whole set of retrofitting recommendations for the specific building based on the building characteristics.

Experiments run and results validation To validate our models, we employed a standard training and testing procedure. The dataset was split into training and test sets, with the training set used to train the models and the test set reserved for validation. The data was split for the experiments in the following proportions:

- 80% of data for training set
- 20% of data for test set

Specifically, we trained each model on the training set and then used the trained models to make predictions on the test set.

Train-test split approach Considering the imbalanced nature of data, we have to pay particular attention to splitting data for training and validation steps. Standard methods for imbalanced data include stratification, which means dividing the dataset into subsets in such a way that each subset maintains a similar distribution of classes as observed in the original dataset. This approach helps to ensure that all classes are adequately represented in each subset, thus preventing bias towards the more frequent classes on training and evaluation steps.

Traditional single-label stratification methods often fail to yield balanced dataset divisions, and this imbalance hinders the ability of classifiers to generalize effectively. Common issues with conventional train/test splits include the complete absence of specific labels in the training set or disproportionately allocating most label pair evidence to the test set. This distribution leaves the training set devoid of adequate evidence necessary for generalizing conditional probabilities associated with label relationships.

Multilabel data stratification addresses these challenges and can be done in several ways, as described in Sechidis, Vlahavas, and Tsoumakas (2011). To handle this issue, we use an iterative stratification mechanism provided by the Scikit-multilearn package, which aims to ensure a more equitable distribution of evidence concerning label relations up to a specified order, which leads to facilitating more effective model training and evaluation (Szymański and Kajdanowicz, 2017).

Classification algorithms We use different algorithms and techniques in multilabel classification experiments. This aimed to identify the most effective approach in predicting a comprehensive set of retrofitting measures based on specific building characteristics.

Linear Model: Logistic Regression

- **MultiOutputClassifier:** Treats each label as an independent binary classification problem.
- **ClassifierChain:** Takes into account possible correlations among labels by chaining them. We evaluated 10 different feature configurations and selected the one that yielded the best performance.

Instance-based Learning: KNeighbors Classifier

- **Multilabel:** A straightforward multilabel approach where each label is treated independently but uses the neighbors' votes to decide.
- **ClassifierChain:** Incorporates label dependencies by chaining predictions. We evaluated 10 different feature configurations and selected the one that yielded the best performance.

Tree-based Methods:

Decision Tree Classifier

- **Multilabel:** Direct approach using decision trees for multilabel classification.
- **MultiOutputClassifier:** Utilizes multiple decision trees, one per label, without considering label dependencies.
- **ClassifierChain:** Decision trees with chained label predictions to model label correlations. We evaluated 10 different feature configurations and selected the one that yielded the best performance.

Random Forest Classifier

- **Multilabel:** Single model for multilabel classification.
- **ClassifierChain:** Uses multiple random forests in a chained fashion to model dependencies among labels. We evaluated 10 different feature configurations and selected the one that yielded the best performance.

Extra Tree Classifier

- **Multilabel:** Single model approach, not explicitly divided into sub-methods for multilabel.
- **ExtraTrees Classifier:** Similar to Random Forest but using Extra Trees for the ensemble, typically leading to more randomized splits.

Ensemble Methods: Random Forest Classifier and ExtraTrees Classifier These methods described above are ensemble methods that leverage multiple trees to improve prediction accuracy and robustness against overfitting.

Neural Network We use a multi-label binary classification neural network that consists of a sequence of four fully connected layers, each followed by a dropout layer with a probability of 0.3, to prevent overfitting. The dimensions of these layers increase progressively (128, 256, 512, 1024), with each layer employing ReLU activation and using BCELoss as a loss function.

5.1.1 Classification on the full dataset

We start with running classifiers on full datasets to identify a baseline for each method. Numerical results, which represented as F1-score and Hamming loss values, are outlined in Table 5.1

TABLE 5.1: Multilabel classification: result summary

Method	Dataset 1		Dataset 2	
	F1-score	Hamming loss	F1-score	Hamming loss
LogisticRegression MultiOutputClassifier	0,0551	0,04	**0,0797**	**0,0388**
LogisticRegression ClassifierChain	0,0553	**0,04**	**0,2436**	0,0486
KNeighborsClassifier Multilabel	0,2298	0,0401	**0,2569**	**0,0392**
KNeighborsClassifier ClassifierChain	0,2859	**0,0449**	**0,3536**	0,0467
DecisionTreeClassifier Multilabel	**0,3825**	**0,0465**	0,374	0,0472
DecisionTreeClassifier MultiOutputClassifier	**0,3533**	**0,0489**	0,3431	0,0499
DecisionTreeClassifier ClassifierChain	**0,3804**	**0,0497**	0,3716	0,0506
RandomForestClassifier	**0,3048**	**0,0299**	0,2879	0,0305

Continued on next page

Continuation of Table 5.1				
Method	Dataset 1		Dataset 2	
	F1-score	Hamming loss	F1-score	Hamming loss
RandomForestClassifier ClassifierChain	**0,4213**	**0,0336**	0,3995	0,0348
ExtraTreeClassifier	**0,3989**	**0,0459**	0,3657	0,0475
ExtraTreesClassifier	**0,3216**	**0,0299**	0,3135	0,0308
Neural Network	**0,1606**	**0,038**	0,1479	0,0385

General Conclusion The multilabel classification experiments demonstrated that ensemble methods, specifically Random Forest and Extra Trees classifiers, in conjunction with the ClassifierChain approach, showed the highest classification performance in terms of F1-score value.

This highlights the promising potential of ensemble methods combined with label dependency modeling to improve prediction accuracy. Specifically, the Random Forest ClassifierChain achieved the highest performance, suggesting that capturing label dependencies might be a promising approach for improving retrofit measure predictions. We need to investigate this deeper with more label configurations in further work.

In contrast, logistic regression models showed relatively low performance, indicating that simpler linear models may struggle with the complexity of multilabel classification in this context and huge amount of features. But with ClassifierChain the Logistic Regression method significantly improved the performance on the second dataset. Dataset 2 contains more binary labels and this performance change may indicate that the combination of the ClassifierChain approach can be applicable for the Logistic Regression as well with less number of numerical features. This direction should be investigated deeply.

The KNeighbors and Decision Tree classifiers performed better, with the Decision Tree showing notable effectiveness in handling the multilabel classification task without the need for label dependency modeling - the ClassifierChain approach did not improve the classification performance.

Despite using the default parameters and not using the hyperparameter tuning and cross-validation technique, the results still provided an understanding of the comparative performance of different methods.

In numbers, the obtained results show a general performance of less than 0.5 F1-score. Despite a Hamming loss value of less than 0.05, this is insufficient to reach the research goal.

5.1.2 Classification on the split data

In this section, we test two splitting approaches to identify if this improves the performance of multilabel classification. Splitting criteria are chosen based on the nature of building stock data.

Spliting by year of construction

In this experiment, we split building stock data based on the year of construction into three groups:

- Built before 1920

- Built between 1920 and 2000
- Built after 2000

The chosen splitting points are driven by the historical nuances of the development of building technology and the housing market in Sweden. This segmentation aimed to explore the variation in multilabel classification performance based on the age of the buildings, using two datasets to evaluate each method's performance.

Summarized results, F1-score for each method per dataset are outlined in Table 5.2. Hamming loss for all options is less than 0.06, which is acceptable, and its difference is not significant to discuss in detail. Full obtained results for each group outlined in Table B.1, Table B.2, Table B.3

TABLE 5.2: Multilabel classification: Classification results on data split by year of construction. F1-score values for different methods and datasets. DS 1 stands for Dataset 1, and DS 2 stands for Dataset 2

Method	Before 1920		1920 - 2000		After 2000	
	DS 1	DS 2	DS 1	DS 2	DS 1	DS 2
LogisticRegression MultiOutputClassifier	0,032	**0,0826**	0,0537	**0,0616**	0,051	**0,2945**
LogisticRegression ClassifierChain	0,0327	**0,2459**	0,0539	**0,2259**	0,0503	**0,4334**
KNeighborsClassifier Multilabel	0,1629	**0,1906**	0,2234	**0,2451**	0,4335	**0,439**
KNeighborsClassifier ClassifierChain	0,2237	**0,3049**	0,2805	**0,3357**	0,4779	**0,5113**
DecisionTreeClassifier Multilabel	**0,3217**	0,3175	**0,3697**	0,3663	**0,5712**	0,5673
DecisionTreeClassifier MultiOutputClassifier	0,2788	**0,2867**	**0,3412**	0,3317	**0,5415**	0,5238
DecisionTreeClassifier ClassifierChain	**0,3183**	0,3168	**0,3687**	0,3606	**0,5731**	0,5649
RandomForestClassifier	**0,2235**	0,2136	**0,2879**	0,2704	**0,5528**	0,5263
RandomForestClassifier ClassifierChain	**0,34**	0,3365	**0,4037**	0,3974	**0,6269**	0,6213
ExtraTreeClassifier	**0,3109**	0,3017	**0,3607**	0,3545	**0,5657**	0,5584
ExtraTreesClassifier	**0,2384**	0,2336	**0,3078**	0,3004	**0,5717**	0,544

General Observations Across all methods, there is a clear trend that more recent construction periods tend to yield better classification results, which can be explained by the nature of the data - most of the buildings in the dataset belong to the two last groups and contain more data for training classifiers. Methods that use the ClassifierChain approach show better results than standard multilabel or MultiOutputClassifier approaches regardless of the group to which data belongs.

Ensemble methods, particularly the Random Forest and Extra Trees classifiers, performed best. The ClassifierChain method significantly improved performance for all groups and methods, similar to classification on the full dataset. DecisionTreeClassifier show similar results with and without ClassifierChain usage.

We can conclude that splitting data improves the prediction performance for the buildings constructed after 2000, and proper data splitting, together with other possible enhancements, may help us to find the building group where the classification approach works well enough to reach the research goal.

Buildings with year of construction earlier than 1920 shows less performance than others, which may be explained by the total amount of building in this group and worse quality of building characteristics for so old building stock. Such building are not standardized the same way as modern buildings and have more individual characteristics.

Splitting by building category

In this experiment, we split building stock data based on the category to which the building belongs into three groups:

- Multi-family dwellings (MFD)
- Office buildings (OB)
- Single- or two-family house (SFH)

Chosen splitting points are driven by differences in building characteristics between these categories. Each category was tested using two datasets, aiming to explore classification performance across different classification methods and building types.

Summarized results, F1-score for each method per dataset are outlined in Table 5.3. Hamming loss for all options is less than 0.07, which is acceptable. Its difference is not significant to discuss in detail. Full obtained results for each group outlined in Table B.4, Table B.5, Table B.6

TABLE 5.3: Multilabel classification: Classification results on data split by building category. F1-score values for different methods and datasets. DS 1 stands for Dataset 1, and DS 2 stands for Dataset 2

Method	MFD		OB		SFH	
	DS 1	DS 2	DS 1	DS 2	DS 1	DS 2
LogisticRegression MultiOutputClassifier	0,0005	**0,0161**	0,0001	**0,0201**	0,131	**0,1701**
LogisticRegression ClassifierChain	0,0005	**0,1086**	0,0001	**0,162**	0,3279	**0,3731**
KNeighborsClassifier Multilabel	0,2279	**0,263**	0,107	**0,129**	0,2626	**0,2853**
KNeighborsClassifier ClassifierChain	0,2682	**0,3201**	0,163	**0,2135**	**0,3545**	0,2853
DecisionTreeClassifier Multilabel	**0,4419**	0,4272	**0,2818**	0,2802	0,3937	**0,3943**
DecisionTreeClassifier MultiOutputClassifier	**0,4271**	0,4086	**0,2723**	0,2617	**0,3574**	0,3518
DecisionTreeClassifier ClassifierChain	**0,4383**	0,4201	**0,2931**	0,2822	**0,3955**	0,3921
RandomForestClassifier	**0,3641**	0,3443	**0,1966**	0,1885	**0,3229**	0,305

Continued on next page

Continuation of Table 5.3						
Method	MFD		OB		SFH	
	DS 1	DS 2	DS 1	DS 2	DS 1	DS 2
RandomForestClassifier ClassifierChain	**0,4051**	0,3956	**0,2995**	0,2889	**0,4832**	0,4597
ExtraTreeClassifier	**0,4531**	0,434	**0,2855**	0,2806	**0,3777**	0,3702
ExtraTreesClassifier	**0,4083**	0,3925	0,2138	**0,2168**	**0,3282**	0,3226

General Observations Across all methods, the family houses category consistently showed higher performance in F1-scores value. Ensemble classifiers provided the best performance among tested. We can conclude that splitting the dataset based on building categories led to improved classification performance compared to using a complete, undivided dataset. This supports the hypothesis that data segmentation based on natural criteria can show better classification performance. Using the ClassifierChain approach led to improving classificaion results.

5.1.3 Conclusion

We experimented with different multilabel classification methods to aim to predict the whole set of recommended retrofitting measures based on the building characteristics. Two dataset versions were used to investigate the performance of classification algorithms on different data characteristics.

Logistic Regression Classifier and KNeighbors Classifier show better performance on Dataset 2, which can be explained by the mostly binary nature of the data in Dataset 2. Tree-based and Ensemble methods show similar results on both dataset versions but slightly better on Dataset 1. The ClassifierChain approach improves the classification results for most experiments, meaning that internal label dependencies are important for the classification on the building stock data. For prediction on the complete dataset, the RandomForest classifier shows the best result with the ClassifierChain approach.

The next stage of experiments is the classification on the split data using two criteria: by year of construction and by building category. We can conclude, that splitting data may lead to prediction performance improvement for most group of data, comparing to the performance of the prediction on the whole dataset. The only group od data, that showed worse classification performance than classification on whole dataset is buildings constructed before 1920, which may be explained by the smaller amount of such data and individual features of this group. For both splitting criteria we got the similar classification performance values.

In summary, we can confirm the promising results for one the our research question "identifying the whole set of retrofitting recommendations for the specific building using multi-label classification methods". Even with default hyperparameters it shows classification performance, enough to continue research in this area.

Another research question was the hypothesis that splitting data based on natural criterion may improve the classification performance, which definitely confirmed by the obtained results.

General prediction performance is less than 90% with an error rate of less than 10% in Hamming loss, which means that we didn't get the good enough performance score to consider the research goal achieved.

5.2 Binary classification

This section outlines experiments with binary classification. We have 28 possible retrofitting measures that may be recommended for a building. In contrast to the multilabel approach, where we tried to predict the whole set of retrofitting recommendations, we consider each measure as an independent binary label here. For every retrofitting measure, we train a separate classifier model on two dataset versions. These experiments aim to test the possibility of predicting a single classification measure with high probability.

Experiments run and results validation We used a standard training and testing procedure to validate the experiments' results. Each dataset was split into training and test sets, with the training set used to train the models and the test set reserved for validation. For all experiments data was split in the following proportions:

- 80% of data for training set
- 20% of data for test set

We trained each model on the training set and then used the trained models to make predictions on the test set and calculate the performance metrics.

Train-test split approach There is a strong imbalance in data for each label, so to mitigate this and ensure proper sampling of train and test data, we use the stratify technique in train_test_split from the sklearn package.

Classification algorithms We use different algorithms in this experiment:

- Linear Model: Logistic Regression
- Instance-based Learning: KNeighbors Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Extra Tree Classifier
- ExtraTrees Classifier

Evaluation The accuracy score does not work well for the imbalanced data since it doesn't capture the positive prediction, but the whole prediction, including negative, and provides false-positive results. We are using F1-score, precision, and recall to identify the performance of each model.

Results We trained classifiers on each version of dataset for every label. Detailed results are outlined in Tables B.7, B.8, B.9, B.10, B.11, B.12. The overall summary for all classifiers and the best classification score is given in Table 5.4.

TABLE 5.4: Binary classification: results summary, best classifier and dataset for each retrofitting measure

	Classifier	Dataset	F1-score	Precision	Recall
R1	ExtraTreesClassifier	DS 1	0,5547	0,9262	0,3959
R2	ExtraTreesClassifier	DS 2	0,6266	0,9326	0,4718
R3	ExtraTreesClassifier	DS 2	0,5828	0,956	0,4191
R4	ExtraTreesClassifier	DS 1	0,6408	0,9729	0,4778
R5	ExtraTreesClassifier	DS 1	0,6318	0,9566	0,4716
R6	ExtraTreesClassifier	DS 2	0,4626	0,8145	0,323
R7	ExtraTreesClassifier	DS 1	0,5093	0,944	0,3487
R8	ExtraTreesClassifier	DS 2	0,5394	0,9631	0,3746
R9	ExtraTreesClassifier	DS 2	0,3949	0,9512	0,2492
R10	ExtraTreesClassifier	DS 2	0,5178	0,9336	0,3582
R11	ExtraTreesClassifier	DS 1	0,5812	0,9519	0,4183
R12	ExtraTreesClassifier	DS 2	0,5018	0,9714	0,3383
R13	ExtraTreesClassifier	DS 1	0,4444	1	0,2857
R14	RandomForest	DS 1	0,6216	0,7878	0,5133
R15	ExtraTreesClassifier	DS 1	0,4054	0,8862	0,2628
R16	ExtraTreesClassifier	DS 2	0,6055	0,9635	0,4415
R17	RandomForest	DS 1	0,6249	0,7789	0,5218
R18	ExtraTreesClassifier	DS 2	0,4817	0,8805	0,3316
R19	ExtraTreesClassifier	DS 2	0,5295	0,9058	0,3741
R20	ExtraTreesClassifier	DS 1	0,5812	0,9345	0,4218
R21	ExtraTreesClassifier	DS 1	0,376	0,6674	0,2617
R22	ExtraTreesClassifier	DS 2	0,6726	0,95	0,5205
R23	ExtraTreesClassifier	DS 2	0,6809	0,9412	0,5333
R24	ExtraTreesClassifier	DS 1	0,5807	0,7866	0,4603
R25	ExtraTreesClassifier	DS 2	0,4031	0,8595	0,2633
R26	ExtraTreesClassifier	DS 1	0,5049	0,9212	0,3478
R27	ExtraTreesClassifier	DS 1	0,3878	0,878	0,2489
R28	ExtraTreesClassifier	DS 1	0,3388	0,8994	0,2087

Ensemble Methods - RandomForest and ExtraTrees show the best performance for the binary classification for all retrofiting measures. Summarised values are provided in Table 5.5.

TABLE 5.5: Single-label classification: aggregated classification performance score

	F1-score	Precision	Recall
Min	0,3388	0,6674	0,2087
Median	0,53445	0,9331	0,37435
Max	0,6809	1	0,5333

Conclusion We tested different binary classification methods aiming to predict the necessity of the specific retrofitting measure recommendation based on building characteristics. Ensemble methods - RandomForest and ExtraTrees classifiers outperform all other classification algorithms. These methods consistently provided high F1-scores, precision, and recall across various retrofit measures. Distinctive

features of the obtained results are high overall precision metric values and significantly small recall metric values, which are caused by data imbalance. Also, despite the high precision value (0.93 in average), all methods showed small enough recall (0.37 in average), which also might be caused by the data imbalance. The average classification performance across all retrofitting measures is 53% in terms of F1-score values, which is not enough to achieve our research goal. Based on the obtained results, we can conclude that the positive validation for our research question, that it is possible to identify a single retrofitting recommendation using classification methods. With default hyperparameters, we got promising results, which confirmed that it makes sense to continue research in this area.

5.3 Experiments limitations discussion

The general number of planned experiments is 6 - three experiments for multilabel classification and three experiments for binary classification approach. We also used two different versions of the dataset in the experiments. From the number of used dataset versions standpoint and computation wise:

- multilabel approach: 12 different classification approaches and 14 different dataset versions - for each version of the dataset, it is one experiment on the full dataset and two experiments with data split into three categories
- single retrofitting measure predictions: for each of 28 retrofitting measures, we run 6 different classification algorithms on two versions of the dataset

This amount of experiments on various datasets led us to the following limitations, which should be considered in the overall results discussion:

Choosing classification algorithms Due to the computational and time-consuming nature of the experiments, we cannot afford to test all possible classification methods. We decided to focus on testing selected methods and left other also promising approaches for future research.

Hyperparameters tunning and cross validation Given the large number of our experiments involving multiple classification approaches and numerous dataset versions, we decided to use default hyperparameters for our models and not do hyperparameter tuning and cross-validation at this stage of the research.

The computational resources and time required for hyperparameter tuning and cross-validation would be extremely high. Hyperparameter tuning typically involves an extensive grid or random search over a range of parameters, exponentially increasing the computational load. Similarly, cross-validation, particularly k-fold cross-validation or its variation for multilabel data stratification, requires training and validating the model multiple times, further compounding the computational demands.

Using default hyperparameters allows us to maintain a feasible and manageable experimental framework while ensuring that our models are trained and evaluated within a reasonable timeframe. Default hyperparameters, although not optimized for each specific dataset, provide a baseline performance that is sufficient for comparative analysis across the numerous classification approaches and dataset versions. This approach is particularly pragmatic given our aim to evaluate the general applicability of various models and methodologies rather than to achieve the absolute best performance for each individual case.

In summary, the decision to use default hyperparameters and avoid extensive tuning and cross-validation is driven by practical considerations of computational efficiency and the necessity to handle a large number of experimental scenarios within the constraints of available resources.

Comparing experiments results The data imbalance and difference in the train-test split processes make it impossible to compare results between experiments using multilabel and binary classification methods. Datasets in multilabel approach experiments are split into train and test subsets, saving the same distribution of all labels simultaneously in train and test. However, for the binary classification experiments, we consider every retrofitting measure as an independent binary and split the datasets with stratification only for this label. This means that it is generally impossible to get the same data distribution in both experiments and comparing the results makes no sense.

Instead, each approach should be evaluated with performance metrics tailored to the specific nature of the classification task at hand. This ensures the evaluation is meaningful and relevant to the respective methodologies.

Chapter 6

Conclusions

6.1 Discussion

In this work, we considered the task of identifying possible retrofitting measures for the building stock based on information about building characteristics and energy demand without using energy consumption modeling tools and models.

We analyzed the related work in the field and outlined the research gaps and formulated the research hypotheses. In our research, we approached the problem of retrofitting recommendations as a classification problem. We explored the Swedish EPC's dataset and created two versions of dataset for experiments, considering provided in EPC data the retrofitting measures recommendations as ground truth set for the classification task. The data is highly imbalanced in terms of number of buildings with- and without recommended retrofitting measures.

In the scope of this research, we investigated different classification approaches and algorithms, including multi- and single-label classification.

Considering data and experiments limitations we should discuss the obtained results for multi- and single-label classification independently.

We can conclude that using multi-label classification methods to identify the whole set of retrofitting recommendations for the specific building is promising but does not show good enough classification performance for using it. Splitting data into groups with similar characteristics improved the classification performance and this approach should be investigated deeply.

Identifying a single retrofitting recommendation using classification methods also provided can be considered a working approach with an average prediction performance of 53%, which is not sufficient to use this method in real-world challenges. But the promising point here is a high level of precision, 0.93 on average, but low recall, 0.37 on average. It means that the binary classification approach can be used for cases when it's more important to be confident in the positive predictions that are made, even if it means missing some actual positive cases. For instance, the enrichment of the building stock datasets with missing retrofitting measures.

Such insufficient classification performance for both tested approaches may be caused by data quality and a high imbalance in data. Also important part here is experiments limitations - further hyperparameter tunning may increase the classification performance.

Concluding the discussion of obtained results from the research questions standpoint, in the scope of this research we confirmed that the classification approach is applicable to choosing possible retrofitting measures based on the building stock data without using the building energy demand evaluation, but requires more efforts in input data validation and additional investigation in proper models' parameters or classification methods.

This thesis marks a significant step in this research area, but there is more work to be done. The methodologies and findings presented in this work can serve as a foundation for future research into retrofitting measures prediction at urban building stock.

6.2 Future work

This work shows that there exists a significant space for improvement in this area. It could be beneficial to investigate approaches to mitigate imbalance in data with adjustments for the specifics of building stock data. For instance, dataset over- and under-sampling techniques should be tested aim to mitigate the data imbalance. Cooperation with industry experts to ensure data quality and the development of a framework for data validation may be crucial for enhancing the accuracy and reliability of predictive models. The data-splitting approach showed performance improvement, and continued investigation of this direction may lead to reaching the goal of recommending retrofitting measures for the specific part of the building stock. Testing the findings on EPC data from other countries can deepen understanding of the domain and the importance of the specific data characteristics and reveal the limitations of the proposed methods. Additionally, further research should consider on the test state-of-the-art classification approaches and hyperparameter tuning to achieve the needed classification performance.

Appendix A

Appendix: dataset structure

A.1 Structure of Swedish EPCs dataset

TABLE A.1: Retrofitting measures recommendations columns in a dataset with artificial recommendation code and recommendation name in English

Code	Column name	Recommendation Name
R1	AtgForslagNyVentil	New radiator valves
R2	AtgForslagJustVarme	Adjustment of heating system
R3	AtgForslagStyrVarme	Time/demand control of heating systems
R4	AtgForslagRengVarme	Cleaning and/or airing of heating systems
R5	AtgForslagBegrTemp	Maximum limitation of internal temperature
R6	AtgForslagNyGivare	New indoor sensor
R7	AtgForslagBytePumpar	Replacement/installation of pressure-controlled pumps
R8	AtgForslagJustVent	Adjustment of ventilation system
R9	AtgForslagTidstyrVent	Time management of ventilation systems
R10	AtgForslagBehovstyrVent	Demand management of ventilation systems
R11	AtgForslagByteFlaktar	Replacement/installation of speed-controlled fans
R12	AtgForslagStyrBelys	Time/demand control of lighting
R13	AtgForslagStyrKyla	Time/demand control of cooling
R14	AtgForslagSparaVatten	Hot water saving measures
R15	AtgForslagEffektivBelys	Energy efficient lighting
R16	AtgForslagIsolKanal	Insulation of pipes and ventilation ducts
R17	AtgForslagByteVarmepump	Replacing/installing a heat pump
R18	AtgForslagByteAnnanVarme	Replacement/installation of a more energy-efficient heat source
R19	AtgForslagByteVent	Replacement/completion of the ventilation system
R20	AtgForslagAterVent	Recovery of ventilation heat
R21	AtgForslagIsolTak	Additional insulation attic joists/roof
R22	AtgForslagIsolVagg	Additional insulation walls

Continuation of Table A.1		
Code	**Column name**	**Recommendation Name**
R23	AtgForslagIsolMark	Additional insulation basement/ground
R24	AtgForslagInstSolceller	Installation of solar cells
R25	AtgForslagInstSolvarme	Installation of solar heating
R26	AtgForslagByteFonster	Change to energy-efficient windows/window doors with inner pane
R27	AtgForslagKompFonster	Complementing windows/window doors with inner pane
R28	AtgForslagTatFonster	Sealing windows/window doors/external doors

TABLE A.2: Data categories encoding

Data category	**Code**
Energy Declaration data	ED
Real estate data	RE
Building information	BI
Building address	BA
Retrofitting measures data	RM

TABLE A.3: Energy declarations dataset structure

Data category code	**Column**	**Description**	**Used for classification**
I. Building Identification			
ED	IdLankod	County code	No
ED	IdLan	County	No
ED	IdKommunkod	Municipal code	No
ED	IdKommun	Municipality	No
RE	IdEgnaHem	Own home	No
RE	IdFastBet	Property designation	No
BI	IdHusnr	House no	No
BI	IdRapportLM	Reason for error report	No
BA	IdAdr	The address of the building	No
BA	IdPostnr	Was built post no	No
BA	IdPostort	The building's postal address	No
BA	IdHuvudadress	Main address	No
II. Building Properties			
ED	EgenTypkod	Type code code	No
ED	EgenTypkod_typ	Type code_type	No
ED	EgenByggnadsKat	Type code_cat	No
ED	EgenKomplexitet	The complexity of the building	Yes

Continued on next page

Continuation of Table A.3

Data category code	Column	Description	Used for classification
ED	EgenByggnadsTyp	Building category	Yes
ED	EgenNybyggAr	Year of new construction	Yes
ED	EgenAtemp	Building area	Yes
ED	EgenAvarmgarage	Dewatering garage	Yes
ED	EgenAntalKallarplan	Number of basement floors heated to >10 degrees	Yes
ED	EgenAntalPlan	Number of floors above ground	Yes
ED	EgenAntalTrapphus	Number of stairwells	Yes
ED	EgenAntalBolgh	Number of residential apartments	Yes
ED	EgenSmaLagenheter	Are there predominantly apartments with a living area of no more than 35 m² each?	Yes
ED	EgenProjVentFlode	Projected average hygienic outdoor air flow in local buildings or apartment buildings	Yes
ED	EgenInstEleffektStorre	There is installed electrical power >10 W/m² for heating and hot water production	Yes
ED	EgenSkyddadEllerVardefull	Is the building protected as a building monument or such a particularly valuable building as referred to in ch. 8 § 13 PBL?	Yes
ED	EgenAtempBostad	Dwellings (incl. secondary area, e.g. stairwell and heated basement)	Yes
ED	EgenAtempHotell	Hotels, boarding houses and student dormitories, %	No
ED	EgenAtempRestaurang	Restaurant, %	No
ED	EgenAtempKontor	Office and administration, %	No
ED	EgenAtempLivsmedel	Store and warehouse space for food, %	No
ED	EgenAtempButik	Shop and warehouse space for other trade, %	No
ED	EgenAtempKopcentrum	Shopping center, %	No
ED	EgenAtempVard	Care, around the clock, %	No
ED	EgenAtempVardDag	Daytime care (including serviced accommodation, hairdressers, etc.), %	No
ED	EgenAtempSkolor	Schools (preschool-university), %	No
ED	EgenAtempBad	Bathing, sports, sports facilities (not outdoor arenas), %	No

Continued on next page

Continuation of Table A.3

Data category code	Column	Description	Used for classification
ED	EgenAtempTeater	Theatre, concert, cinema venues and other gathering spaces, %	No
ED	EgenAtempOvrigaVad	Other activities - specify what	No
ED	EgenAtempOvrig	Other activities, %	No
ED	EgenAtempSumma	Total business	Yes
III. Energy use			
ED	EgiForstaArManad	The data refer to the first month of the 12-month period	Yes
ED	EgiSistaArManad	The last month of the 12-month period the data refers to	Yes
ED	EgiBeraknatVarde	Estimated energy use	No
ED	EgiFjarrvarmeUPPV	Energy for heating - District heating (1)	Yes
ED	EgiFjarrvarmeVV	Energy for domestic hot water - District heating (1)	Yes
ED	EgiOljaUPPV	Energy for heating - Heating oil (2)	Yes
ED	EgiOljaVV	Energy for domestic hot water - Heating oil (2)	Yes
ED	EgiGasUPPV	Energy for heating - Natural gas, city gas (3)	Yes
ED	EgiGasVV	Energy for domestic hot water - Natural gas, city gas (3)	Yes
ED	EgiVedUPPV	Energy for heating - Firewood (4)	Yes
ED	EgiVedVV	Energy for domestic hot water - Wood (4)	Yes
ED	EgiFlisUPPV	Energy for heating - Chips/pellets/briquettes (5)	Yes
ED	EgiFlisVV	Energy for domestic hot water - Chips/pellets/briquettes (5)	Yes
ED	EgiOvrBiobransleUPPV	Energy for heating - Other biofuel (6)	Yes
ED	EgiOvrBiobransleVV	Energy for domestic hot water - Other biofuel (6)	Yes
ED	EgiElVattenUPPV	Energy for heating - Electricity (water borne) (7)	Yes
ED	EgiElDirektUPPV	Energy for heating - Electricity (direct acting) (8)	Yes

Continued on next page

Continuation of Table A.3

Data category code	Column	Description	Used for classification
ED	EgiElLuftUPPV	Energy for heating - Electricity (air borne) (9)	Yes
ED	EgiPumpMarkUPPV	Energy for heating - Ground source heat pump (electric) (10)	Yes
ED	EgiPumpFranluftUPPV	Energy for heating - Heat pump exhaust air (electricity) (11)	Yes
ED	EgiPumpLuftLuftUPPV	Energy for heating - Heat pump-air/air (electricity) (12)	Yes
ED	EgiPumpLuftVattenUPPV	Energy for heating - Heat pump-air/water (electricity) (13)	Yes
ED	EgiElVV	Energy for heating - Domestic hot water (electricity) (14)	Yes
ED	EgiFjarrkyla	District cooling (15)	Yes
ED	EgiKomfort	Electricity for comfort cooling (16)	Yes
ED	EgiFastighet	Property electricity (17)	Yes
ED	EgiHushall	Household (18)	No
ED	EgiVerksamhet	Business electricity (19)	No
ED	EgiSumma2	Total 1-17	No
ED	EgiSolvarme	Enter solar collector area	Yes
ED	EgiGruppSolvarme	Is there solar heating?	Yes
ED	EgiBerEngProduktion	Estimated energy production	No
ED	EgiStationEI	City (Energy Index)	No
ED	EgiSolcell	Enter solar cell area	Yes
ED	EgiGruppSolcell	Are there solar cell systems?	Yes
ED	EgiBerElProduktion	Estimated electricity production	No
ED	EgiEnergianvandning	The building's energy use (Normal annual corrected value (Energy index))	Yes
ED	EgiPrimarenergianvandning	The building's primary energy use	Yes
ED	EgiPrimarenergital2019	Used to calculate Energy Performance from 2019-01-01 to 2020-08-31 Available in the appendix from 2020-09-01	No
ED	EgiPrimarenergital2020	Used to calculate Energy Performance from 2020-09-01 -> Available on the appendix from 2020-09-01	No

Continued on next page

Continuation of Table A.3

Data category code	Column	Description	Used for classification
ED	EgiSpecifikEnergianvandning	Used to Energy performance from 2010 to 2018-12-31 Available on attachment from 2020-09-01	No
ED	EgiRefvarde1	Reference value 1 (according to new construction requirements)	No
ED	EgiRefvarde2Max	Reference value 2 (similar buildings)	No
ED	EgiRefvarde3	Reference value 3 (new construction requirements for this building)	No
ED	EgiEnergiPrestanda	Energy performance	No
ED	EgiVersion	Energy class calculated on reference value year	No
ED	EgiEnergiklass	Energy class	No
IV. Information about radon			
ED	RadGruppHaltMatt	Is the radon level measured?	Yes
ED	RadHalt	Radon level	Yes
ED	RadTypMatning	Type of measurement	No
ED	RadMatDatum	Date of radon measurement	No
V. Information on ventilation control			
ED	VentGruppKrav	Are there requirements for regular ventilation checks in the building?	Yes
ED	VentTypFTX	FTX	Yes
ED	VentTypF	F	Yes
ED	VentTypFT	FT	Yes
ED	VentTypSjalvdrag	Self trait	Yes
ED	VentTypFmed	F with recycling	Yes
ED	VentGruppGodkand	Is the ventilation check carried out at the time of the energy declaration?	No
ED	VentDelvisProcent	% without remark	No
ED	VentGruppUtanAnm	Is the ventilation control unremarked at the time of the energy declaration?	No
VI. Air conditioning system details			
ED	LuftGruppFinnsluft	There are air conditioning systems with a nominal cooling output greater than 12 kW	Yes
VII. Inspection of heating systems			

Continued on next page

Continuation of Table A.3

Data category code	Column	Description	Used for classification
ED	InspUppvGruppNomStorre	Is there a heating system or combined space heating and ventilation system with a nominal space heating output of more than 70 kW?	Yes
ED	InspUppvBedomningNomEffekt	Assessment basis for determining nominal power	No
ED	InspUppvGruppInspSkyldighet	Is the building subject to inspection obligations?	No
ED	InspUppvAngeNomEffekt	Enter the nominal power of the system	No
ED	InspUppvAngeNomEffektYta	Enter area served	No
ED	InspUppvGruppLamplig	Is the size and efficiency of the air conditioning system appropriate for the needs of the building?	No
ED	InspUppvLampligKommentar	Comment	No
ED	InspUppvUndAvtalEgipres	Agreement on energy performance (section 8 a first paragraph 1 EDF)	No
ED	InspUppvUndSysFastStyr	System for property automation/property management (Section 8 a first paragraph 2 EDF)	No
ED	InspUppvUndFunkOverReglBost	Function for monitoring and regulation, residential buildings (section 8 a first paragraph 3 EDF)	No
VIII. Air conditioning system inspection			
ED	InspLuftGruppNomStorre	Is there an air conditioning system or combined air conditioning and ventilation system with a rated output of more than 70 kW?	Yes
ED	InspLuftBedomningNomEffekt	Assessment basis for determining nominal power	No
ED	InspLuftGruppInspSkyldighet	Is the building subject to inspection obligations?	No
ED	InspLuftAngeNomEffekt	Enter the nominal power of the system	No
ED	InspLuftAngeNomEffektYta	Enter area served	No
ED	InspLuftGruppLamplig	Is the size and efficiency of the air conditioning system appropriate for the needs of the building?	No
ED	InspLuftLampligKommentar	Comment	No

Continued on next page

Continuation of Table A.3

Data category code	Column	Description	Used for classification
ED	InspLuftUndAvtalEgipres	Agreement on energy performance (section 8 a first paragraph 1 EDF)	No
ED	InspLuftUndSysFastStyr	System for property automation/property management (Section 8 a first paragraph 2 EDF)	No
ED	InspLuftUndFunkOverReglBost	Function for monitoring and regulation, residential buildings (section 8 a first paragraph 3 EDF)	No
IX. Carried out energy efficiency measures since the previous energy declaration			
RM	AtgUtfordaNyVentil	New radiator valves	Yes
RM	AtgUtfordaJustVarme	Adjustment of heating system	Yes
RM	AtgUtfordaStyrVarme	Time/demand control of heating systems	Yes
RM	AtgUtfordaRengVarme	Cleaning and/or airing of heating systems	Yes
RM	AtgUtfordaBegrTemp	Maximum limitation of internal temperature	Yes
RM	AtgUtfordaNyGivare	New indoor sensor	Yes
RM	AtgUtfordaBytePumpar	Replacement/installation of pressure-controlled pumps	Yes
RM	AtgUtfordaAnnanVarme	Other action (heat)	No
RM	AtgUtfordaJustVent	Adjustment of ventilation system	Yes
RM	AtgUtfordaTidstyrVent	Time management of ventilation systems	Yes
RM	AtgUtfordaBehovstyrVent	Demand management of ventilation systems	Yes
RM	AtgUtfordaByteFlaktar	Replacement/installation of speed-controlled fans	Yes
RM	AtgUtfordaAnnanVent	Other measure (ventilation)	No
RM	AtgUtfordaStyrBelys	Time/demand control of lighting	Yes
RM	AtgUtfordaStyrKyla	Time/demand control of cooling	Yes
RM	AtgUtfordaAnnanBelysKyla	Other measures (lighting, cooling, etc.)	No
RM	AtgUtfordaInstSolceller	Installation of solar cells	Yes
RM	AtgUtfordaInstSolvarme	Installation of solar heating	Yes
RM	AtgUtfordaIsolTak	Additional insulation attic joists/roof	Yes
RM	AtgUtfordaIsolVagg	Additional insulation walls	Yes

Continued on next page

Continuation of Table A.3

Data category code	Column	Description	Used for classification
RM	AtgUtfordaIsolMark	Additional insulation basement/ground	Yes
RM	AtgUtfordaByteFonster	Change to energy-efficient windows/window doors with inner pane	Yes
RM	AtgUtfordaKompFonster	Complementing windows/window doors with inner pane	Yes
RM	AtgUtfordaTatFonster	Sealing windows/window doors/external doors	Yes
RM	AtgUtfordaAnnanBygg	Other action (build)	No
RM	AtgUtfordaSparaVatten	Hot water saving measures	Yes
RM	AtgUtfordaEffektivBelys	Energy efficient lighting	Yes
RM	AtgUtfordaIsolKanal	Insulation of pipes and ventilation ducts	Yes
RM	AtgUtfordaByteVarmepump	Replacing/installing a heat pump	Yes
RM	AtgUtfordaByteAnnanVarme	Replacement/installation of a more energy-efficient heat source	Yes
RM	AtgUtfordaByteVent	Replacement/completion of the ventilation system	Yes
RM	AtgUtfordaAterVent	Recovery of ventilation heat	Yes
RM	AtgUtfordaAnnanInst	Other action (installation)	No
RM	AtgUtfordaUtfortAr	Performed (year)	No
RM	AtgUtfordaBeskrivning	Description of the action	No
X. Recommendations on cost-effective measures			
RM	AtgForslagNyVentil	New radiator valves	Yes
RM	AtgForslagJustVarme	Adjustment of heating system	Yes
RM	AtgForslagStyrVarme	Time/demand control of heating systems	Yes
RM	AtgForslagRengVarme	Cleaning and/or airing of heating systems	Yes
RM	AtgForslagBegrTemp	Maximum limitation of internal temperature	Yes
RM	AtgForslagNyGivare	New indoor sensor	Yes
RM	AtgForslagBytePumpar	Replacement/installation of pressure-controlled pumps	Yes
RM	AtgForslagAnnanVarme	Other action (heat)	No
RM	AtgForslagJustVent	Adjustment of ventilation system	Yes
RM	AtgForslagTidstyrVent	Time management of ventilation systems	Yes
RM	AtgForslagBehovstyrVent	Demand management of ventilation systems	Yes

Continued on next page

Continuation of Table A.3

Data category code	Column	Description	Used for classification
RM	AtgForslagByteFlaktar	Replacement/installation of speed-controlled fans	Yes
RM	AtgForslagAnnanVent	Other measure (ventilation)	No
RM	AtgForslagStyrBelys	Time/demand control of lighting	Yes
RM	AtgForslagStyrKyla	Time/demand control of cooling	Yes
RM	AtgForslagAnnanBelysKyla	Other measures (lighting, cooling, etc.)	No
RM	AtgForslagSparaVatten	Hot water saving measures	Yes
RM	AtgForslagEffektivBelys	Energy efficient lighting	Yes
RM	AtgForslagIsolKanal	Insulation of pipes and ventilation ducts	Yes
RM	AtgForslagByteVarmepump	Replacing/installing a heat pump	Yes
RM	AtgForslagByteAnnanVarme	Replacement/installation of a more energy-efficient heat source	Yes
RM	AtgForslagByteVent	Replacement/completion of the ventilation system	Yes
RM	AtgForslagAterVent	Recovery of ventilation heat	Yes
RM	AtgForslagAnnanInst	Other action (installation)	No
RM	AtgForslagIsolTak	Additional insulation attic joists/roof	Yes
RM	AtgForslagIsolVagg	Additional insulation walls	Yes
RM	AtgForslagIsolMark	Additional insulation basement/ground	Yes
RM	AtgForslagInstSolceller	Installation of solar cells	Yes
RM	AtgForslagInstSolvarme	Installation of solar heating	Yes
RM	AtgForslagByteFonster	Change to energy-efficient windows/window doors with inner pane	Yes
RM	AtgForslagKompFonster	Complementing windows/window doors with inner pane	Yes
RM	AtgForslagTatFonster	Sealing windows/window doors/external doors	Yes
RM	AtgForslagAnnanBygg	Other action (build)	No
RM	AtgForslagEgiMinskad	Reduced energy use	No
RM	AtgForslagKostnad	Cost per kWh saved	No
ED	ExpertGruppHosUppdragsgivare	Are you employed by the person who is obliged to ensure that there is an energy declaration or an inspection report?	No
ED	Godkand	Approval	No

Continued on next page

Continuation of Table A.3			
Data category code	**Column**	**Description**	**Used for classification**
ED	Version	Version	No
XI. Miscellaneous			
ED	OvrBesiktigat	Has the building been inspected on site?	No
ED	OvrBesiktigatUndantag	If no, which exception is invoked	No
XII. Expert			
ED	ExpertGodkand	Date of approval	No
ED	ExpertBehorighet	Expert qualification	No
XIII. Building's energy performance			
ED	EgiSpecifikEnergianvandning	Specific energy use according to BBR 24 and earlier	No
ED	EgiPrimarenergital2019	Primary energy number according to BBR 25	No
ED	EgiPrimarenergital2020	Primary energy number according to BBR 29	No

A.2 Statistics of classification dataset

The table below contains numeric statistics for the created classification dataset columns, excluded all retrofitting measures columns. Total number of rows is 274878.

TABLE A.4: Statistics of classification dataset

Column	Data type	Missed values	Skewness	Min	Q50	Max
EgenKomplexitet	Categorical	0	-	-	-	-
EgenByggnadsTyp	Categorical	0	-	-	-	-
EgenNybyggAr	Numeric	0	-4,34	1006	1969	2023
year_bucket	Numeric	0	-4,99	1000	1960	2020
EgenAtemp	Numeric	0	14,95	50	230	209514
EgenAvarmgarage	Numeric	174501	39,85	0	0	65992
EgenAntalKallarplan	Numeric	55456	14,62	0	0	20
EgenAntalPlan	Numeric	61734	5,25	1	2	96
EgenAntalTrapphus	Numeric	75551	8,77	0	1	72
EgenAntalBolgh	Numeric	74652	12,01	0	2	1059

Continued on next page

Continuation of Table A.4

Column	Data type	Missed values	Skewness	Min	Q50	Max
EgenSmaLagenheter	Binary	0	-	-	-	-
EgenProjVentFlode	Numeric	0	4,73	0	0	9,65
EgenInstEleffektStorre	Binary	0	-	-	-	-
EgenSkyddadEllerVardefull	Categorical	70223	-	-	-	-
EgenAtempBostad	Numeric	0	-2,11	0	100	100
EgenAtempBusiness	Numeric	0	2,11	0	0	100
EgiFjarrvarmeUPPV	Numeric	151957	9,3	0	93606	16249577
EgiFjarrvarmeVV	Numeric	154265	11,74	0	15000	2535456
EgiOljaUPPV	Numeric	270005	9,2	0	16000	1815806
EgiOljaVV	Numeric	271606	10,03	0	3200	199512
EgiGasUPPV	Numeric	271951	4,08	0	20433	1722749
EgiGasVV	Numeric	272182	5,29	0	4889	338200
EgiVedUPPV	Numeric	248415	4,62	0	2500	206890
EgiVedVV	Numeric	263338	2,85	0	0	38150
EgiFlisUPPV	Numeric	267395	10,64	0	20705	2891000
EgiFlisVV	Numeric	268356	17,08	0	3958	548700
EgiOvrBiobransleUPPV	Numeric	268162	13,04	0	0	2175600
EgiOvrBiobransleVV	Numeric	268247	12,95	0	0	198000
EgiElVattenUPPV	Numeric	250385	25,68	0	9718	2905000
EgiElDirektUPPV	Numeric	200897	17,54	0	4215	1516380
EgiElLuftUPPV	Numeric	264480	9,56	0	5684,5	1042512
EgiPumpMarkUPPV	Numeric	239758	15,1	0	9600	2840000
EgiPumpFranluftUPPV	Numeric	253735	8,39	0	6791	590765
EgiPumpLuftLuftUPPV	Numeric	234616	23,22	0	3882,5	470000
EgiPumpLuftVattenUPPV	Numeric	260310	12,03	0	8200	809085
EgiElVV	Numeric	116019	18,38	0	2154	466898
EgiFjarrkyla	Numeric	265480	13,42	0	0	5523306
EgiKomfort	Numeric	256561	12,03	0	300	2354650
EgiFastighet	Numeric	63072	32,47	0	2242	10257000
EgiSolvarme	Numeric	272223	21,52	0	12	3000

Continued on next page

Continuation of Table A.4

Column	**Data type**	**Missed values**	**Skewness**	**Min**	**Q50**	**Max**
EgiGruppSolvarme	Binary	0	-	-	-	-
EgiSolcell	Numeric	269733	5,6	0	68	7000
EgiGruppSolcell	Binary	0	-	-	-	-
EgiEnergianvandning	Numeric	0	15,62	213	25384	26748348
EgiPrimarenergianvandning	Numeric	0	18,56	383	32592	29981704
VentGruppKrav	Binary	0	-	-	-	-
VentTypFTX	Binary	0	-	-	-	-
VentTypF	Binary	0	-	-	-	-
VentTypFT	Binary	0	-	-	-	-
VentTypSjalvdrag	Binary	0	-	-	-	-
VentTypFmed	Binary	0	-	-	-	-
LuftGruppFinnsluft	Binary	0	-	-	-	-
InspUppvGruppNomStorre	Binary	0	-	-	-	-
InspLuftGruppNomStorre	Binary	0	-	-	-	-

Appendix B

Appendix: Results of experiments

B.1 Multilabel classification results on split data

B.1.1 Classification results on data split by year of construction

This section contains a results summary for multilabel classification for data split on the building year of construction.

TABLE B.1: Multilabel classification: results' summary for building built before 1920

Method	Dataset 1		Dataset 2	
	F1-score	Hamming loss	F1-score	Hamming loss
LogisticRegression MultiOutputClassifier	0,032	0,0393	0,0826	0,0385
LogisticRegression ClassifierChain	0,0327	0,0392	0,2459	0,0527
KNeighborsClassifier Multilabel	0,1629	0,0418	0,1906	0,0411
KNeighborsClassifier ClassifierChain	0,2237	0,0449	0,3049	0,0495
DecisionTreeClassifier Multilabel	0,3217	0,053	0,3175	0,0536
DecisionTreeClassifier MultiOutputClassifier	0,2788	0,0567	0,2867	0,0559
DecisionTreeClassifier ClassifierChain	0,3183	0,0559	0,3168	0,0573
RandomForestClassifier	0,2235	0,0333	0,2136	0,0335
RandomForestClassifier ClassifierChain	0,34	0,0384	0,3365	0,0388
ExtraTreeClassifier	0,3109	0,0533	0,3017	0,0537
ExtraTreesClassifier	0,2384	0,0336	0,2336	0,0346

TABLE B.2: Multilabel classification: results' summary for building built between 1920 and 2000

Method	Dataset 1		Dataset 2	
	F1-score	Hamming loss	F1-score	Hamming loss
LogisticRegression MultiOutputClassifier	0,0537	0,0403	0,0616	0,0392
LogisticRegression ClassifierChain	0,0539	0,0402	0,2259	0,0543
KNeighborsClassifier Multilabel	0,2234	0,0405	0,2451	0,0396
KNeighborsClassifier ClassifierChain	0,2805	0,0451	0,3357	0,0476
DecisionTreeClassifier Multilabel	0,3697	0,0472	0,3663	0,0478
DecisionTreeClassifier MultiOutputClassifier	0,3412	0,0503	0,3317	0,0506
DecisionTreeClassifier ClassifierChain	0,3687	0,0504	0,3606	0,0517
RandomForestClassifier	0,2879	0,0305	0,2704	0,0311
RandomForestClassifier ClassifierChain	0,4037	0,0348	0,3974	0,0364
ExtraTreeClassifier	0,3607	0,0478	0,3545	0,0481
ExtraTreesClassifier	0,3078	0,0306	0,3004	0,0314

TABLE B.3: Multilabel classification: results' summary for building built after 2000

Method	Dataset 1		Dataset 2	
	F1-score	Hamming loss	F1-score	Hamming loss
LogisticRegression MultiOutputClassifier	0,051	0,0389	0,2945	0,0344
LogisticRegression ClassifierChain	0,0503	0,0388	0,4334	0,0412
KNeighborsClassifier Multilabel	0,4335	0,0322	0,439	0,0322
KNeighborsClassifier ClassifierChain	0,4779	0,034	0,5113	0,0354
DecisionTreeClassifier Multilabel	0,5712	0,0322	0,5673	0,0321
DecisionTreeClassifier MultiOutputClassifier	0,5415	0,0336	0,5238	0,0345
DecisionTreeClassifier ClassifierChain	0,5731	0,0341	0,5649	0,0336
RandomForestClassifier	0,5528	0,0209	0,5263	0,0219
RandomForestClassifier ClassifierChain	0,6269	0,0235	0,6213	0,0246
ExtraTreeClassifier	0,5657	0,0323	0,5584	0,0328
ExtraTreesClassifier	0,5717	0,0213	0,544	0,022

B.1.2 Classification results on data split by building category

This section contains a results summary for multilabel classification for data split on the building's category.

TABLE B.4: Multilabel classification: results' summary for building category "Multi-family dwellings"

Method	Dataset 1		Dataset 2	
	F1-score	Hamming loss	F1-score	Hamming loss
LogisticRegression MultiOutputClassifier	0,0005	0,0422	0,0161	0,0421
LogisticRegression ClassifierChain	0,0005	0,0422	0,1086	0,0512
KNeighborsClassifier Multilabel	0,2279	0,0409	0,263	0,0398
KNeighborsClassifier ClassifierChain	0,2682	0,0469	0,3201	0,0488
DecisionTreeClassifier Multilabel	0,4419	0,042	0,4272	0,0432
DecisionTreeClassifier MultiOutputClassifier	0,4271	0,0449	0,4086	0,0468
DecisionTreeClassifier ClassifierChain	0,4383	0,0471	0,4201	0,0493
RandomForestClassifier	0,3641	0,026	0,3443	0,0269
RandomForestClassifier ClassifierChain	0,4051	0,0277	0,3956	0,029
ExtraTreeClassifier	0,4531	0,0409	0,434	0,0429
ExtraTreesClassifier	0,4083	0,0244	0,3925	0,0253

TABLE B.5: Multilabel classification: results' summary for building category "Office buildings"

Method	Dataset 1		Dataset 2	
	F1-score	Hamming loss	F1-score	Hamming loss
LogisticRegression MultiOutputClassifier	0,0001	0,0435	0,0201	0,0433
LogisticRegression ClassifierChain	0,0001	0,0435	0,162	0,0519
KNeighborsClassifier Multilabel	0,107	0,0468	0,129	0,0465
KNeighborsClassifier ClassifierChain	0,163	0,056	0,2135	0,0595
DecisionTreeClassifier Multilabel	0,2818	0,0607	0,2802	0,061
DecisionTreeClassifier MultiOutputClassifier	0,2723	0,0652	0,2617	0,0663

Continued on next page

Continuation of Table B.5				
Method	Dataset 1		Dataset 2	
	F1-score	Hamming loss	F1-score	Hamming loss
DecisionTreeClassifier ClassifierChain	0,2931	0,0674	0,2822	0,0675
RandomForestClassifier	0,1966	0,0351	0,1885	0,035
RandomForestClassifier ClassifierChain	0,2995	0,0426	0,2889	0,0425
ExtraTreeClassifier	0,2855	0,0605	0,2806	0,0602
ExtraTreesClassifier	0,2138	0,035	0,2168	0,0347

TABLE B.6: Multilabel classification: results' summary for building category "Single- or two-family house"

Method	Dataset 1		Dataset 2	
	F1-score	Hamming loss	F1-score	Hamming loss
LogisticRegression MultiOutputClassifier	0,131	0,0357	0,1701	0,0351
LogisticRegression ClassifierChain	0,3279	0,0438	0,3731	0,0447
KNeighborsClassifier Multilabel	0,2626	0,0373	0,2853	0,0366
KNeighborsClassifier ClassifierChain	0,3545	0,0419	0,2853	0,0366
DecisionTreeClassifier Multilabel	0,3937	0,0445	0,3943	0,0444
DecisionTreeClassifier MultiOutputClassifier	0,3574	0,0462	0,3518	0,0462
DecisionTreeClassifier ClassifierChain	0,3955	0,0458	0,3921	0,0466
RandomForestClassifier	0,3229	0,0297	0,305	0,03
RandomForestClassifier ClassifierChain	0,4832	0,0349	0,4597	0,0348
ExtraTreeClassifier	0,3777	0,0455	0,3702	0,0461
ExtraTreesClassifier	0,3282	0,0303	0,3226	0,031

B.2 Single-label classification results

TABLE B.7: Single-label classification: results' summary for LogisticRegression Classifier

	Dataset 1			Dataset 2		
	F1-score	Precision	Recall	F1-score	Precision	Recall
R1	0	0	0	0	0	0
R2	0	0	0	0	0	0
R3	0	0	0	0,0149	0,6923	0,0075
R4	0	0	0	0	0	0

Continued on next page

Continuation of Table B.7						
	Dataset 1			Dataset 2		
	F1-score	Precision	Recall	F1-score	Precision	Recall
R5	0	0	0	0,0038	0,4545	0,0019
R6	0	0	0	0,0062	0,3889	0,0031
R7	0	0	0	0	0	0
R8	0	0	0	0	0	0
R9	0	0	0	0	0	0
R10	0	0	0	0	0	0
R11	0	0	0	0	0	0
R12	0	0	0	0	0	0
R13	0	0	0	0	0	0
R14	0,023	0,1437	0,0125	0,1092	0,6177	0,0599
R15	0	0	0	0	0	0
R16	0	0	0	0	0	0
R17	0,4179	0,5101	0,354	0,422	0,5966	0,3265
R18	0	0	0	0,0012	0,25	0,0006
R19	0	0	0	0,0018	0,1429	0,0009
R20	0	0	0	0,0019	0,1111	0,001
R21	0,0262	0,2506	0,0138	0,0084	0,5686	0,0042
R22	0	0	0	0	0	0
R23	0	0	0	0	0	0
R24	0	0	0	0,1964	0,5899	0,1178
R25	0	0	0	0	0	0
R26	0	0	0	0	0	0
R27	0,0022	1	0,0011	0,0047	0,25	0,0024
R28	0	0	0	0	0	0

TABLE B.8: Single-label classification: results' summary for KNeighbors Classifier

	Dataset 1			Dataset 2		
	F1-score	Precision	Recall	F1-score	Precision	Recall
R1	0,3458	0,5788	0,2465	0,3849	0,5843	0,287
R2	0,4268	0,5757	0,3391	0,4904	0,628	0,4023
R3	0,3331	0,6658	0,2221	0,4135	0,7102	0,2917
R4	0,3543	0,6948	0,2378	0,3944	0,6382	0,2854
R5	0,3868	0,6504	0,2753	0,4308	0,6343	0,3261
R6	0,3267	0,5419	0,2339	0,3763	0,5752	0,2796
R7	0,2709	0,6605	0,1704	0,2864	0,6196	0,1863
R8	0,2815	0,6105	0,1829	0,3426	0,6615	0,2312
R9	0,163	0,44	0,1	0,1891	0,5	0,1166
R10	0,2236	0,5057	0,1436	0,326	0,6848	0,2139
R11	0,2842	0,5625	0,1901	0,3773	0,6327	0,2688
R12	0,248	0,6739	0,152	0,2126	0,5094	0,1343
R13	0,0444	0,3333	0,0238	0,16	0,4444	0,0976
R14	0,4469	0,553	0,3749	0,4993	0,6081	0,4235
R15	0,1951	0,56	0,1181	0,2774	0,5832	0,182

Continued on next page

Continuation of Table B.8						
	Dataset 1			Dataset 2		
	F1-score	Precision	Recall	F1-score	Precision	Recall
R16	0,2968	0,604	0,1968	0,3679	0,624	0,2609
R17	0,5224	0,6047	0,4598	0,5424	0,6135	0,4861
R18	0,2679	0,5856	0,1737	0,3444	0,6266	0,2374
R19	0,3117	0,5548	0,2168	0,382	0,6454	0,2713
R20	0,311	0,6158	0,208	0,3987	0,636	0,2903
R21	0,2966	0,4395	0,2239	0,3025	0,4443	0,2293
R22	0,4198	0,7391	0,2931	0,5106	0,7636	0,3836
R23	0,4812	0,8421	0,3368	0,5344	0,8537	0,3889
R24	0,4272	0,5363	0,355	0,4789	0,5738	0,411
R25	0,2149	0,6118	0,1303	0,2994	0,7075	0,1899
R26	0,2819	0,6257	0,182	0,338	0,6217	0,2321
R27	0,1635	0,5513	0,096	0,1972	0,4839	0,1238
R28	0,1715	0,529	0,1023	0,1924	0,4924	0,1195

TABLE B.9: Single-label classification: results' summary for Decision-Tree Classifier

	Dataset 1			Dataset 2		
	F1-score	Precision	Recall	F1-score	Precision	Recall
R1	0,4562	0,4387	0,4751	0,4412	0,459	0,283
R2	0,5343	0,5216	0,5476	0,5364	0,5505	0,3665
R3	0,411	0,3997	0,423	0,4159	0,435	0,2625
R4	0,4557	0,425	0,4911	0,443	0,4449	0,2845
R5	0,4834	0,4668	0,5013	0,4616	0,4882	0,3
R6	0,4025	0,3915	0,4142	0,4111	0,4293	0,2587
R7	0,3456	0,3324	0,3599	0,3079	0,3235	0,182
R8	0,3632	0,3588	0,3676	0,3686	0,3996	0,2259
R9	0,3049	0,2966	0,3136	0,3153	0,345	0,1872
R10	0,372	0,3396	0,4111	0,3739	0,4041	0,23
R11	0,4168	0,3952	0,4408	0,4129	0,4436	0,2602
R12	0,2844	0,2602	0,3137	0,2993	0,3134	0,176
R13	0,2885	0,2419	0,3571	0,2418	0,2683	0,1375
R14	0,5346	0,5271	0,5424	0,5253	0,5367	0,3562
R15	0,3122	0,3037	0,3212	0,3098	0,3272	0,1833
R16	0,3752	0,3699	0,3806	0,4025	0,4314	0,252
R17	0,5479	0,5442	0,5517	0,5411	0,5532	0,3709
R18	0,3642	0,3465	0,3837	0,3705	0,4014	0,2274
R19	0,3925	0,3855	0,3996	0,4054	0,4287	0,2543
R20	0,428	0,4158	0,4408	0,4094	0,436	0,2574
R21	0,3689	0,3584	0,38	0,3612	0,3757	0,2204
R22	0,4008	0,3745	0,431	0,4431	0,4977	0,2846
R23	0,45	0,4286	0,4737	0,5028	0,5	0,3358
R24	0,5155	0,5072	0,5241	0,5139	0,5297	0,3458
R25	0,2556	0,2422	0,2707	0,2515	0,2709	0,1438
R26	0,3729	0,3542	0,3937	0,3636	0,3831	0,2222
R27	0,2737	0,2556	0,2946	0,2311	0,2453	0,1307

Continuation of Table B.9						
	Dataset 1			Dataset 2		
	F1-score	Precision	Recall	F1-score	Precision	Recall
R28	0,2634	0,249	0,2796	0,2413	0,2584	0,1372

TABLE B.10: Single-label classification: results' summary for RandomForest Classifier

	Dataset 1			Dataset 2		
	F1-score	Precision	Recall	F1-score	Precision	Recall
R1	0,5202	0,9682	0,3556	0,4644	0,97	0,3053
R2	0,5969	0,9699	0,4311	0,5859	0,9723	0,4192
R3	0,5379	0,9869	0,3697	0,5268	0,9907	0,3588
R4	0,5701	0,9784	0,4022	0,5577	0,9721	0,391
R5	0,6004	0,9829	0,4322	0,5737	0,9771	0,406
R6	0,4252	0,8786	0,2805	0,4296	0,9034	0,2818
R7	0,4604	0,9793	0,301	0,4231	0,9821	0,2696
R8	0,4681	0,9888	0,3066	0,502	0,9843	0,3369
R9	0,3686	0,974	0,2273	0,3803	1	0,2348
R10	0,4761	0,9604	0,3165	0,5069	0,9712	0,343
R11	0,5403	0,9779	0,3732	0,5315	0,9731	0,3656
R12	0,4047	0,9811	0,2549	0,4479	1	0,2886
R13	0,4444	1	0,2857	0,3922	1	0,2439
R14	0,6216	0,7878	0,5133	0,6054	0,7938	0,4894
R15	0,3717	0,9395	0,2317	0,343	0,9438	0,2096
R16	0,5278	0,9344	0,3677	0,564	0,9675	0,398
R17	0,6249	0,7789	0,5218	0,6059	0,7721	0,4985
R18	0,4501	0,9568	0,2943	0,4515	0,9506	0,296
R19	0,484	0,9504	0,3247	0,5044	0,9561	0,3426
R20	0,5237	0,9766	0,3578	0,5193	0,9604	0,3558
R21	0,349	0,7291	0,2294	0,3222	0,7071	0,2087
R22	0,5775	0,9794	0,4095	0,6626	0,9909	0,4977
R23	0,6573	0,9792	0,4947	0,6466	1	0,4778
R24	0,5727	0,8166	0,441	0,5597	0,8055	0,4288
R25	0,3681	1	0,2256	0,3595	0,9775	0,2203
R26	0,4663	0,9783	0,3061	0,4617	0,9913	0,301
R27	0,3339	0,9381	0,2031	0,2903	0,9603	0,171
R28	0,3092	0,967	0,184	0,2754	0,9537	0,1609

TABLE B.11: Single-label classification: results' summary for Extra-Tree Classifier

	Dataset 1			Dataset 2		
	F1-score	Precision	Recall	F1-score	Precision	Recall
R1	0,4564	0,45	0,463	0,4412	0,4399	0,4424
R2	0,5396	0,5398	0,5395	0,5395	0,5307	0,5486
R3	0,4382	0,4331	0,4434	0,4214	0,4148	0,4283
R4	0,4756	0,4756	0,4756	0,4321	0,4122	0,4539

Continued on next page

Continuation of Table B.11						
	Dataset 1			Dataset 2		
	F1-score	Precision	Recall	F1-score	Precision	Recall
R5	0,5054	0,4946	0,5167	0,4793	0,4642	0,4954
R6	0,3937	0,389	0,3985	0,4001	0,3908	0,4098
R7	0,3714	0,3627	0,3806	0,3557	0,3474	0,3644
R8	0,369	0,3687	0,3693	0,3872	0,374	0,4014
R9	0,3156	0,3087	0,3227	0,3206	0,3139	0,3275
R10	0,3751	0,3547	0,398	0,3846	0,3726	0,3973
R11	0,4342	0,4041	0,469	0,4479	0,4368	0,4595
R12	0,3529	0,3529	0,3529	0,344	0,3398	0,3483
R13	0,425	0,4474	0,4048	0,3023	0,2889	0,3171
R14	0,524	0,527	0,5211	0,5034	0,4999	0,5069
R15	0,308	0,2998	0,3167	0,3055	0,3003	0,3108
R16	0,4045	0,4058	0,4032	0,4251	0,4099	0,4415
R17	0,535	0,5343	0,5358	0,5315	0,5267	0,5363
R18	0,3724	0,3642	0,3808	0,3881	0,3747	0,4026
R19	0,4166	0,4222	0,4112	0,4041	0,3915	0,4176
R20	0,4484	0,4492	0,4475	0,446	0,4325	0,4604
R21	0,3689	0,3646	0,3734	0,3654	0,3604	0,3706
R22	0,4167	0,4032	0,431	0,4882	0,4597	0,5205
R23	0,4828	0,4537	0,5158	0,5134	0,4948	0,5333
R24	0,5089	0,5048	0,5131	0,5001	0,491	0,5094
R25	0,245	0,2444	0,2456	0,2801	0,2721	0,2886
R26	0,3722	0,3669	0,3776	0,3733	0,3672	0,3795
R27	0,2687	0,2591	0,279	0,2534	0,2456	0,2618
R28	0,2696	0,2668	0,2724	0,2371	0,2333	0,2409

TABLE B.12: Single-label classification: results' summary for Extra-Trees Classifier

	Dataset 1			Dataset 2		
	F1-score	Precision	Recall	F1-score	Precision	Recall
R1	0,5547	0,9262	0,3959	0,5239	0,9011	0,3693
R2	0,6191	0,9281	0,4645	0,6266	0,9326	0,4718
R3	0,5721	0,9559	0,4082	0,5828	0,956	0,4191
R4	0,6408	0,9729	0,4778	0,6398	0,9554	0,4809
R5	0,6318	0,9566	0,4716	0,6165	0,9486	0,4566
R6	0,4497	0,821	0,3097	0,4626	0,8145	0,323
R7	0,5093	0,944	0,3487	0,4958	0,9283	0,3382
R8	0,5052	0,9703	0,3415	0,5394	0,9631	0,3746
R9	0,3771	0,9568	0,2348	0,3949	0,9512	0,2492
R10	0,5066	0,9464	0,3458	0,5178	0,9336	0,3582
R11	0,5812	0,9519	0,4183	0,5691	0,9589	0,4046
R12	0,4797	0,9701	0,3186	0,5018	0,9714	0,3383
R13	0,4444	1	0,2857	0,3922	1	0,2439
R14	0,6198	0,7726	0,5175	0,6026	0,7649	0,4972
R15	0,4054	0,8862	0,2628	0,3974	0,8408	0,2602

Continued on next page

Continuation of Table B.12						
	Dataset 1			Dataset 2		
	F1-score	Precision	Recall	F1-score	Precision	Recall
R16	0,5475	0,9167	0,3903	0,6055	0,9635	0,4415
R17	0,619	0,7543	0,5248	0,6172	0,7266	0,5364
R18	0,4723	0,9038	0,3197	0,4817	0,8805	0,3316
R19	0,5145	0,9217	0,3568	0,5295	0,9058	0,3741
R20	0,5812	0,9345	0,4218	0,5646	0,9284	0,4057
R21	0,376	0,6674	0,2617	0,3637	0,6343	0,2549
R22	0,5946	0,9802	0,4267	0,6726	0,95	0,5205
R23	0,6712	0,9608	0,5158	0,6809	0,9412	0,5333
R24	0,5807	0,7866	0,4603	0,5726	0,7578	0,4602
R25	0,3796	0,8661	0,2431	0,4031	0,8595	0,2633
R26	0,5049	0,9212	0,3478	0,4977	0,9502	0,3372
R27	0,3878	0,878	0,2489	0,3444	0,8916	0,2134
R28	0,3388	0,8994	0,2087	0,3054	0,8306	0,1871

Bibliography

Ali, Usman et al. (2019). "A data-driven approach for multi-scale building archetypes development". In: *Energy and Buildings*.

Ali, Usman et al. (2020). "A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings". In: *Applied Energy*.

Ali, Usman et al. (2024). "Urban building energy performance prediction and retrofit analysis using data-driven machine learning approach". In: *Energy and Buildings*.

Alrobaie, Abdurahman and Moncef Krarti (2022). "A Review of Data-Driven Approaches for Measurement and Verification Analysis of Building Energy Retrofits". In: *Energies*.

Biessmann, Felix, Bhaskar Kamble, and Rita Streblow (2023). "An Automated Machine Learning Approach towards Energy Saving Estimates in Public Buildings". In: *Energies*.

BPIE (2019). http://bpie.eu. The Buildings Performance Institute Europe (BPIE).

Dahlström, Lukas, Tor Broström, and Joakim Widén (2022). "Advancing urban building energy modelling through new model components and applications: A review". In: *Energy & Buildings*.

Deb, C. and A. Schlueter (2021). "Review of data-driven energy modelling techniques for building retrofit". In: *Renewable and Sustainable Energy Reviews*.

DIRECTIVE 2002/91/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL (2002). https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:001:0065:0071:en:PDF.

DIRECTIVE 2010/31/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL (2010). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02010L0031-20210101&rid=5.

ENTRANZE (2019). http://www.entranze.eu. ENTRANZE-ENforce the TRAnsition to Nearly Zero energy buildings, Intelligent Energy Europe programme.

Eurostat (2019). http://ec.europa.eu/eurostat. Eurostat by Statistical office of the European Union.

Feng, Kailun et al. (2022). "Energy-Efficient Retrofitting under Incomplete Information: A Data-Driven Approach and Empirical Study of Sweden". In: *Buildings*.

Ferrando, Martina et al. (2020). "Urban building energy modeling (UBEM) tools: A state-of-the-art review of bottom-up physics-based approaches". In: *Sustainable Cities and Society*.

Ferrantelli, Andrea and Jarek Kurnitski (2022). "Energy Performance Certificate Classes Rating Methods Tested with Data: How Does the Application of Minimum Energy Performance Standards to Worst-Performing Buildings Affect Renovation Rates, Costs, Emissions, Energy Consumption?" In: *Energies*.

Gibaja, E. and S. Ventura (2015). "A tutorial on multilabel learning". In: *ACM Computing Surveys (CSUR)*.

Grillone, Benedetto et al. (2020). "A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings". In: *Renewable and Sustainable Energy Reviews*.

In-depth follow-up municipal energy and climate advice (EKR) (2022). www.technopolis-group.com/.

Johari, Fatemeh, Farshid Shadram, and Joakim Widén (2023). "Urban building energy modeling from geo-referenced energy performance certificate data: Development, calibration, and validation". In: *Sustainable cities and society*.

Ma, Zhenjun et al. (2012). "Existing building retrofits: Methodology and state-of-the-art". In: *Energy and Buildings*.

Marasco, Daniel E. and Constantine E. Kontokosta (2016). "Applications of machine learning methods to identifying and predicting building retrofit opportunities". In: *Energy and Buildings*.

ODYSSEE-MURE (2019). http://www.odyssee-mure.eu. ODYSSEE and MURE data bases.

Pan, Yiqun et al. (2023). "Building energy simulation and its application for building performance optimization: A review of methods, tools, and case studies". In: *Advances in Applied Energy*.

Pasichnyi, Oleksii et al. (2019). "Energy performance certificates — New opportunities for data-enabled urban energy policy instruments?" In: *Energy Policy*.

Pedone, Livio et al. (2023). "Energy refurbishment planning of Italian school buildings using data-driven predictive models". In: *Applied Energy*.

Sechidis, Konstantinos, I. Vlahavas, and Grigorios Tsoumakas (2011). "On the Stratification of Multi-label Data". In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Seyedzadeha, Saleh et al. (2020). "Machine learning modelling for predicting non-domestic buildings energy performance: a model to support deep energy retrofit decision-making". In: *Applied Energy*.

Shu, Lei and Dong Zhao (2023). "Decision-Making Approach to Urban Energy Retrofit—A Comprehensive Review". In: *Buildings*.

Szymański, P. and T. Kajdanowicz (Feb. 2017). "A scikit-based Python environment for performing multi-label classification". In: *ArXiv e-prints*. arXiv: 1702.01460 [cs.LG].

TABULA (2019). http://episcope.eu. TABULA-typology approach for building stock energy assessment.

The Paris Agreement (2015). https://www.un.org/en/climatechange/paris-agreement.

Thrampoulidis, Emmanouil, Gabriela Hug, and Kristina Orehounig (2023). "Approximating optimal building retrofit solutions for large-scale retrofit analysis". In: *Applied Energy*.

Thrampoulidis, Emmanouil et al. (2021). "A machine learning-based surrogate model to approximate optimal building retrofit solutions". In: *Applied Energy*.

Tsoumakas, G., I. Katakis, and I. Vlahavas (2010). "Mining multi-label data". In.

Zhang, Haonan et al. (2022). "Artificial Neural Network for Predicting Building Energy Performance: A Surrogate Energy Retrofits Decision Support Framework". In: *Buildings*.