Koutsandreas, Diamantis; Keppo, Ilkka

Harnessing machine learning algorithms to unveil energy efficiency investment archetypes

Published in:
Energy Reports

DOI:
10.1016/j.egyr.2024.09.009

Published: 01/12/2024

Document Version
Publisher's PDF, also known as Version of record

Energy Reports 12 (2024) 3180–3195

Contents lists available at ScienceDirect

Energy Reports

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

Harnessing machine learning algorithms to unveil energy efficiency investment archetypes

Diamantis Koutsandreas [a,b,*], Ilkka Keppo [a]

[a] Department of Mechanical Engineering, Aalto University, Otakaari 4, 02150 Espoo, Finland
[b] International Institute for Applied Systems Analysis, Schlossplatz 1, 2361 Laxenburg, Austria

ARTICLE INFO

Keywords:
Clustering analysis
Investment ensembles
Unsupervised learning
Supervised learning
Classification models

ABSTRACT

Increasing transparency about the performance of different projects is crucial to reducing the heterogeneity in the energy efficiency services market, thereby upscaling investments. In this context, machine learning algorithms could assist in identifying and analyzing energy efficiency project archetypes, although this field has so far been explored with a limited view in the literature. This paper aims to address this gap by identifying energy efficiency investment families and the determinant factors of the classification scheme, using machine learning. In this effort, it hinges on a wide range of indicators from implemented projects around Europe and the USA, including investment profitability, initial investment, risk of failure, intervention type, life measure, region of implementation and building type. The analysis employs two clustering approaches, namely Partitioning Around Medoids (PAM) and K-means, determining the number of clusters based on the Silhouette index and total within–cluster sum of squares. The results indicate that energy efficiency investments can be classified into three categories: (i) "junk investments", characterized by low–profitability (IRR~10%), moderate risk, and extended horizons; (ii) "safe profitability", distinguished by high profitability (IRR~30%) and minimal risk; and (iii) "high stakes", described by exceptionally high profitability (IRR~40%), coupled with a substantial risk. Next to profitability and risk of failure, also energy efficiency intervention and building type (sector) emerge among the most influential factors in the classification scheme. Feature importance shows a significant sensitivity to the chosen classification model.

1. Introduction

Boosting energy efficiency is widely recognized as crucial for achieving the established energy and climate targets (Rubino, 2017). However, apart from the behavioral factors (e.g., rebound effect) that hinder progress in this area (Sorrell et al., 2020), the current pace of mobilizing public and private capital for energy efficiency falls significantly short of what is needed to keep these targets within reach (Deloitte, 2016; IEA, 2021). This shortfall can be primarily ascribed to the heterogeneity of the energy efficiency services market. This variety arises from the interplay of energy efficiency projects with a multitude of technical aspects (e.g., quality of technology, experience of involved workers) and economic factors (e.g., energy prices, economic environment). Addressing this complexity necessitates resource-intensive, bottom-up approaches for effective treatment and evaluation (Hill, 2019; Stevens et al., 2019). Furthermore, the lack of available data on successful projects exacerbates the challenge (Bremer et al., 2024), complicating the benchmarking of investments.

Consequently, investors face a transparency gap when estimating the returns and associated risks of an energy efficiency project (Koutsandreas et al., 2022; Mexis et al., 2021). These risks reflect the potential for payment default, where capital providers may not recover their funds (Rezessy and Bertoldi, 2010). Due to the challenges in identifying and accurately assessing the risks of energy efficiency investments, financing decisions are often solely based on the borrower's creditworthiness. This approach can lead investors and financial institutions to overlook many projects with significant potential (Wang et al., 2017), which may not be recognized as such without thorough analysis. Additionally, investors typically seek a deep understanding of the sectors where they invest, thereby avoiding these types of investments due to perceived uncertainties and complexities. Notably, the risk of future losses often weighs more heavily on investors' decisions than the potential for gains (Wang et al., 2017).

Addressing the above-mentioned challenges in energy efficiency financing requires standardized decision support methods and models (Kleanthis et al., 2022), which have been the focus of several studies

* Corresponding author at: Department of Mechanical Engineering, Aalto University, Otakaari 4, 02150 Espoo, Finland.
E-mail address: diamantis.koutsandreas@aalto.fi (D. Koutsandreas).

https://doi.org/10.1016/j.egyr.2024.09.009
Received 4 January 2024; Received in revised form 23 August 2024; Accepted 4 September 2024
Available online 11 September 2024

over the last years. Indicatively, Doukas et al. (2021) employed a combination of traditional and machine learning classification techniques to predict the performance of energy efficiency investments. Similarly, Sarmas et al. (2022) applied machine learning classification methods, in conjunction with a meta-learning model, to determine the optimal allocation of energy efficiency investment portfolios, according to renovation costs and energy savings. Additionally, Kleanthis et al. (2022) introduced an assessment framework for energy efficiency investment profitability across various investor profiles. Relatedly, Mexis et al. (2021) developed a methodology based on the multicriteria method ELECTRE Tri to classify energy efficiency investments into predetermined classes, each defined by specific limits to each indicator (e.g., profitability).

The existing body of literature has predominantly focused on predicting and classifying the performance of energy efficiency investments using historical data from successfully implemented projects, utilizing a type of supervised learning or predefined classification rules. Supervised learning involves the division of available data into input and output: training models to accurately project output data (e.g., profitability) when fed with the input data. Such type of algorithms has been also extensively utilized for various applications in the broader energy sector (e.g., Chen et al. (2023), Li et al. (2024), Lahmiri (2024) and Al Kez et al. (2024)). However, there remains a gap in the literature regarding the examination of the relationships between diverse energy efficiency projects with a view to understanding the coexistence of certain project characteristics (e.g., country of implementation, associated interventions, or technical characteristics) and how these interact with project profitability and risk. Analyzing these aspects based on which projects resemble or differ can help reduce the heterogeneity in the energy efficiency services market, thereby identifying project ensembles that might not have been considered a priori. This can in turn support investors in making informed decisions. Addressing this requires the application of an unsupervised machine-learning approach to a multidimensional dataset of completed projects. Unlike supervised learning, where models are informed about the target variable to be predicted, unsupervised learning models seek relevance and affinity across different unspecified elements based on various features. From this type of models, clustering is of particular interest, as it can systematically categorize elements to maximize intra-group similarities and minimize inter-group similarities (Ahmad and Khan, 2019), thereby efficiently formulating ensembles. Additionally, it facilitates the study of group characteristics, thereby providing insights into the influential factors for categorization. Therefore, clustering holds significant potential to address the identified literature gap.

Clustering techniques have been applied within a limited scope in the context of energy efficiency, primarily focusing on specific sectors (e.g., buildings or industry) and countries, or/and a narrow set of factors (e.g., building energy intensity). For instance, Liao and He (2018) employed a clustering approach to classify 37 industrial subsectors in China into energy efficiency levels, seeking for the influential factors of their energy performance. Similarly, Papadopoulos et al. (2018) categorized commercial and housing buildings in New York City based on energy intensity, exploring the influential factors of their temporal energy demand. In another study, Geyer et al. (2017) grouped buildings according to the cost savings resulting from various retrofit measures.

However, clustering techniques have been extensively applied within the broader energy sector. Indicatively, they have been employed to investigate representative groups within highly heterogeneous datasets, such as decarbonization pathways for the UK (Li et al., 2020; Pye et al., 2019), land types (Shivakumar et al., 2021), and potential sites for ocean renewable energy (Uti et al., 2023). These techniques have also been instrumental in identifying the key determinant factors for a member's inclusion in a specific group, as seen in studies on transformation strategy ensembles in the power sector (Moksnes et al., 2019). Notably, these studies have focused on attributes that can predict successful projects (e.g., profitability), overlooking indicators of potentially unsuccessful projects like the level of risk. However, the latter aspects can be crucial for involved actors in the energy efficiency services market (Doukas, 2018), whose returns are linked to the project's successful implementation. Therefore, a comprehensive evaluation of these projects should include such indicators as well.

The paper contributes to the literature by examining how various project characteristics coexist in energy efficiency investments and the implications of these factors for project success or failure. Its primary innovation lies in the consideration of a multitude of attributes of such projects across several regions, while also addressing overlooked factors associated with project failure (e.g., risk). Based on these factors, the paper explores the classification of these investments into key ensembles, based on combinations of these attributes. Specifically, the classification hinges on a wide range of indicators from completed projects in Europe and the USA, including investment profitability, initial investment, annual savings, risk of failure, intervention type, life measure, region of implementation, sector, and building type. The initial dataset is sourced from the De–risking Energy Efficiency Platform (DEEP) (EEFIG, 2017), whereas further modifications and enhancements are implemented to facilitate an efficient clustering. These mainly include projects' regional classification, risk assessment, outliers' identification, and analysis of feature correlations.

From a methodological point of view, this study contributes to the existing literature legacy by comparing two clustering approaches to uncover hidden patterns within the analyzed data. The first, Partitioning Around Medoids (PAM) (Botyarov and Miller, 2022), treats the dataset as mixed and is applied upon a custom distance matrix. The other, K-means (Uti et al., 2023), handles the dataset as purely numeric, converting all categorical variables via one-hot encoding (see below). The optimal number of clusters is determined by the Silhouette index (Asri et al., 2019) and the total within-cluster sum of squares (WCSS) (Brusco and Steinley, 2007). The study tries to identify the key determinant factors within this classification system to understand the critical aspects than may move one project from one class to another. In this regard, a Random Forest model (Metzig et al., 2020) is trained to predict the cluster to which energy efficiency projects belong, in turn calculating the average decrease in the Gini index (Bouke et al., 2023) across model's decision trees. Moreover, the study explores the sensitivity of results to the chosen classification model by applying additional key classification models, including Gradient Boosting, Support Vector Machines, and Logistic Regression.

The analysis outcomes can fundamentally assist in reducing heterogeneity in the energy efficiency services market, thereby enhancing the understanding of involved actors for such investments. The value of this research is underscored by the existing transparency gap in these investments (Loureiro et al., 2020). Investors typically seek to diversify their portfolios across various projects in several regions and sectors, considering aspects like risk and profitability. Identifying archetypes of energy efficiency investments beyond single-factor classification schemes (e.g., profitability-based) can assist investors in building well-diversified portfolios across multiple sectors and regions, while also increasing invested capital. This approach enables investors to better understand how the risk and profitability of their portfolios are formulated when adding projects from different countries, sectors, and interventions, thereby making them feel more secure in leveraging further capital in this investment type.

Additionally, the analysis outcomes equip stakeholders with insights into the interplay of project characteristics, which can in turn facilitate the prediction of unknown project attributes and the assessment of their impact. Furthermore, this information can assist policymakers in formulating more efficient and coherent energy efficiency policies. Indicatively, they can articulate support mechanisms not just at an intervention level, but also for sets of interventions that share similar characteristics. Similarly, from a European perspective, policymakers can develop stimulus packages tailored to groups of similar countries.

This can, in turn, reduce the heterogeneity of energy efficiency policies, thereby making the relevant policy framework more efficient and coherent.

The remainder of the paper is organized as follows. Section 2 outlines the experimental design the study adopts to address the posed research questions, including the description of the dataset utilized and the preprocessing tasks applied to it. Following, Section 3 presents the key results arising from the implementation of the adopted framework to the provided data, while Section 4 concludes the paper and suggests potential avenues for future research.

2. Experimental design

2.1. Input data and preprocessing tasks

The analysis commences by utilizing the DEEP database as the primary data source. A set of modifications and additions are then implemented to it to facilitate an effective clustering. The analysis first seeks to eliminate the significant outliers in the IRR of projects, as this is one of the fundamental practices of efficient clustering. While these outliers may be due to poor data collection, they empirically confirm the heterogeneity that exists in the energy efficiency services market (see Introduction for more details). Due to the non-normal distribution of these IRR values across projects, the analysis avoids an approach designed for normal distribution, such as the Z-score treatment of outliers. Instead, it employs interquartile ranges to identify and remove outliers (Yaro et al., 2023). Specifically, a project is considered an outlier if its IRR exceeds or falls below 1.5 times the interquartile range, defined as the difference between the third quartile (Q3) and the first quartile (Q1). This widely accepted threshold of 1.5 is chosen because it effectively eliminates extreme values while retaining an important part of the initial dataset, thereby ensuring the representativeness of the remaining data to the original dataset. Consequently, the remaining project values fall within the bounds set by Eq. (1). This method eliminates projects that significantly diverge from the rest of the dataset, thereby enhancing the consistency and robustness of the input data.

$$Q1 - 1.5 \times (Q3 - Q1) \leq \text{(eligible IRR values)} \leq Q3 + 1.5 \times (Q3 - Q1) \quad (1)$$

The application of this criterion leads to the identification and subsequent removal of 756 project outliers, resulting in a refined dataset of 5456 successfully implemented projects. Fig. 1 displays boxplots representing the IRR distribution for the remaining projects following the exclusion of outliers, including the minimum, first quartile, median, third quartile, and maximum IRR values within the dataset. IRR distributions are categorized according to the energy efficiency measures and sectors.

One of the fundamental prerequisites for efficient clustering analysis is ensuring that each nominal attribute has neither too few nor too many levels, and that there is a sufficient sample size for each category without significant asymmetries (Ghattas et al., 2017). In light of these, the analysis undertakes a set of preprocessing tasks on the dataset's nominal variables. First, regarding regional classification of available projects in the original dataset, there are huge disparities among European countries, with some having only a few projects. Consequently, clustering this dataset with the original regional classification could compromise the efficiency of the process. Nonetheless, incorporating regional information in the clustering procedure can provide valuable insights for the analysis.

To address the asymmetry of available data across countries, the study reclassifies them into regional coalitions based on the United Nations Geoscheme (Shvili, 2021), whereas the USA remains a distinct regional coalition. The United Nations Geoscheme classifies European countries into "Eastern Europe", "Western Europe", "Northern Europe", and "Southern Europe". This scheme is chosen as it provides a robust, widely-accepted regional classification framework primarily based on geographical proximity, while also factoring in cultural, economic, and historical factors. Consequently, it ensures homogeneity within each region, such as common weather conditions that can significantly affect the performance and risk of energy efficiency projects. The resulting distribution of energy efficiency projects across these regions is as follows: (i) Western Europe: 3975; (ii) Southern Europe: 46; (iii) Eastern Europe: 940; (iv) Northern Europe: 687; (v) USA: 558.

The next step in the prepossessing tasks involves the recategorization of building types included in the original dataset. This is done to reduce project heterogeneity by grouping similar building types together, thereby reducing the excessive number of levels in this attribute. Specifically, first, all types of "family buildings" (e.g., single- or multi-family buildings) are grouped under the "Households" category. Similarly, the various types of public buildings (e.g., education buildings) are classified as "Public" buildings. For the several building types pertaining to the service sector, the analysis groups them into the "Trade and Services" category, while also merging the "Street lighting" energy efficiency measure into the "Lighting" category.

Furthermore, the analysis assesses the risk of failure for energy efficiency projects in the dataset. Apart from the factors indicating project success, what is equally important from an investor's perspective is the likelihood that a project will not perform as expected. In this context, "risk of failure" refers to the probability that a project will not meet its predicted performance due to internal factors (e.g., unsuitability of involved workers) or external factors (e.g., energy prices), multiplied by the ratio to which the actual performance may differ from the projected performance.

The quantification of the risk of the projects included in the dataset is performed based on the method proposed by Kleanthis et al. (2022). This method evaluates the technical attributes of the associated energy efficiency measures, specifically focusing on the rebound effect and technical complexity. For each project, based on the involved intervention, the analysis assigns two risk values: one for the rebound effect and another for technical complexity, in turn averaging them to calculate the project's total risk. These risk values are selected over the following scales: "Insignificant"-0, "Low"-0.25, "Medium"-0.5, "High"-1 for rebound effect; and "Low"-0, "Medium"-0.5, "High"-1 for technical complexity. Consequently, total project risk lies in the [0,1] range, where 0 indicates an almost risk-free investment and 1 denotes highly risky ones. The risk values considered for the examined interventions regarding rebound effect and technical complexity are visualized in Fig. 2. These risks are then incorporated into the dataset as an additional feature. However, the analysis does not consider the risk pertaining to the country of implementation (e.g., economic environment or energy price volatility), as it regards wider regional coalitions instead of individual countries. Similarly, the analysis does not factor in the risk stemming from the bottom-up characteristics of the projects (e.g., the experience of technical workers) due to the unavailability of this information.

Table 1 provides an overview of the dataset's features resulting from the aforementioned modifications and additions, summarizing also the values for each variable. Each feature encapsulates a particular aspect of successfully implemented energy efficiency projects, be it profitability, risk of failure, or a particular technical attribute. In cases where specific feature data is unavailable, the analysis assigns the descriptor "unknown". This approach allows for retaining features with missing data, which can inform the clustering process — dropping these features would significantly reduce the available dataset. The sectors of the analyzed projects include buildings and industry. The buildings sector includes energy efficiency interventions designed to reduce energy consumption in buildings, influenced by the needs and actions of the occupants, such as heating and cooling. On the other hand, the industry sector comprises interventions aimed at improving the energy efficiency of industrial processes.

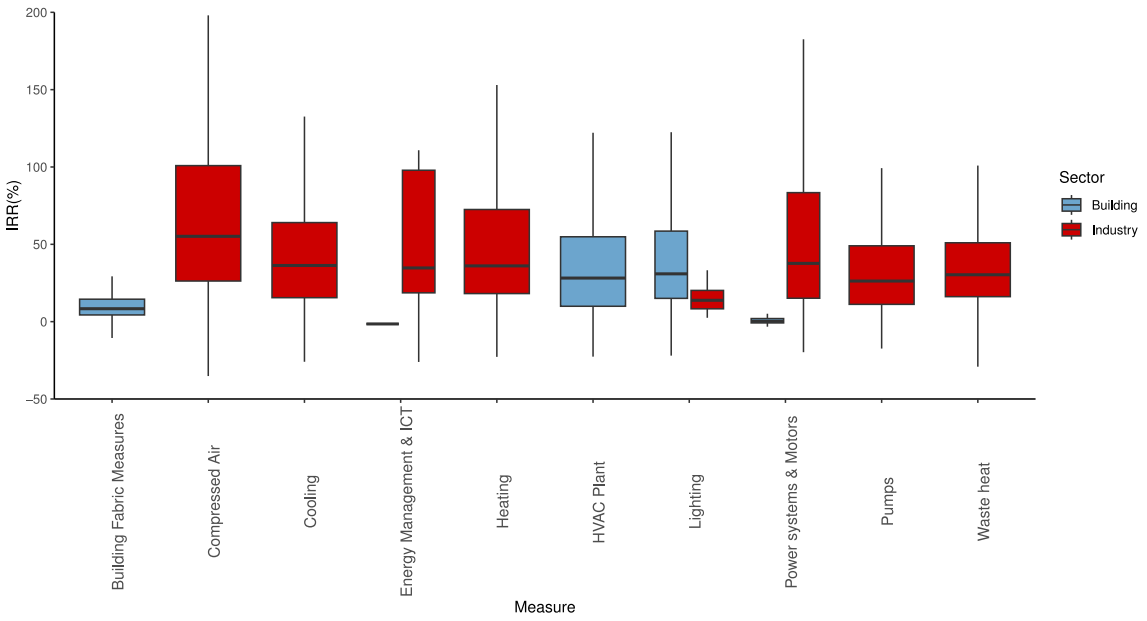

Fig. 1. Boxplots illustrating the range (i.e., minimum and maximum) and quartile values of the project Internal Rate of Return (IRR) for the analyzed energy efficiency investments across sectors and measures.

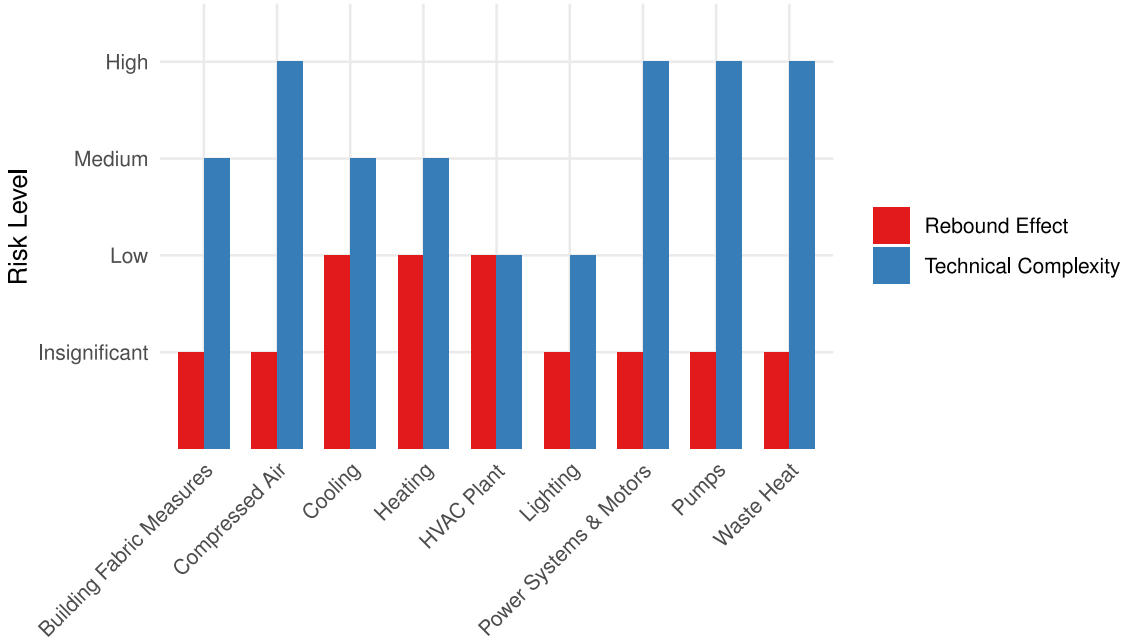

Fig. 2. Assigned risk values for rebound effect and technical complexity across analyzed energy efficiency interventions (Kleanthis et al., 2022).

2.2. Clustering approach

2.2.1. Overview

There are two general approaches for clustering datasets with mixed data types. One approach involves converting all data to numeric form before applying clustering algorithms; the other clusters the data as is, using metrics that can handle both numeric and categorical variables (Ienco et al., 2012). Given the lack of consensus in the literature on which approach is preferable, the analysis employs both methods to cluster the analyzed energy efficiency projects. Following, the study assesses the sensitivity of the results to the chosen clustering approach by comparing the variability between the outcomes of these two methods.

At the first, the dataset is treated as mixed without any conversion of its categorical variables. In this case, the Partitioning Around Medoids (PAM) method (Botyarov and Miller, 2022) is utilized since it allows for handling a custom dissimilarity matrix, as the one arising from using the Gower Dissimilarity (GD) metric (Belenguer et al., 2023) in the examined dataset. In the second approach, the categorical variables of the dataset are converted into numerical ones via one-hot encoding (Hastie et al., 2009), in turn applying the K-Means method (Liu et al., 2023).

In both cases, the first step is to calculate and visualize the correlations between the dataset's features to detect and exclude highly correlated features. Pearson's correlation coefficients (Jebli et al., 2021) are calculated for the numerical variables, while Cramer's V values (Babu and Gajanan, 2022), derived from chi-square tests, are computed for

Table 1
Overview of dataset features utilized to uncover families of energy efficiency investments. In case of data unavailability for certain features, the descriptor "unknown" is assigned.

Variable	Unit	Description	Summary
Project IRR	%	Project IRR considering 100% equity funds with no debt	Numeric value: Min. = −35%, 1st Qu. = 9%, Median = 24%, Mean = 38%, 3rd Qu. = 53%, Max. = 186%
Sector	–	Sector to which each project belongs	Linguistic values: "Building" "Industry"
Initial investment	Thousand €	Initial investment for energy efficiency purposes	Numeric value: Min. = 0.1, 1st Qu. = 500, Median = 2351, Mean = 13743, 3rd Qu. = 8110, Max. = 1 840 000
Annual monetary saving	Thousand €	Annual savings resulting from energy efficiency measures	Numeric value: Min. = 0.01, 1st Qu. = 1.71, Median = 5.38, Mean = 31.94, 3rd Qu. = 14.39, Max. = 4740.74
Life measure	Years	Life of the implemented energy efficiency measures	Numeric value: Min. = 4, 1st Qu. = 12, Median = 12, Mean = 16, 3rd Qu. = 15, Max. = 30
Measure	–	Interventions implemented in each project	Linguistic values: "Energy Management & ICT", "Waste heat (without power generation)", "Cooling", "Pumps", "Power systems & Motors", "Compressed Air", "HVAC Plant", "Heating", "Building Fabric Measures", "Other", "Lighting"
Region	–	Region of project implementation	Linguistic values: "Eastern Europe", "Northern Europe", "Western Europe", "Southern Europe", "USA"
Sub-sector	–	Sub-sector to which each project belongs	Linguistic values: "Administrative and support services", "Electricity, gas, steam and air conditioning supply", "Water/waste management", "Professional, scientific and technical", "Public administration and defence", "Arts, entertainment and recreation", "Mining/quarrying", "Construction", "Education", "Transportation and storage", "Agriculture/forestry/fishing", "Wholesale and retail trade/motor vehicles", "Human health and social work", "Accommodation and food service", "Information and communication", "Other", "Manufacturing", "Finance/insurance", "Real estate", "Unknown"
Organization size	–	Size of the organization where retrofits are applied	Linguistic values: "Unknown", "SMALL", "LARGE", "MEDIUM", "MICRO"
Buildingtype	–	Type of building where retrofits are applied	Linguistic values: "Public", "Households", "Industry", "Trade and Services", "Unknown"
Riskvalue	–	Risk of failure value for each project	Numeric value: Min. = 0, 1st Qu. = 0.125, Median = 0.25, Mean = 0.26, 3rd Qu. = 0.5, Max. = 0.5 (a risk value of 0 corresponds to risk-free investments; a risk value of 1 denotes highly risky investments)

the categorical features. All visualizations are performed with the "ggplot2" package in R (Wickham, 2016).

2.2.2. Mixed data clustering

This stage commences by quantifying the dissimilarities among dataset's energy efficiency projects. This is performed using the GD metric (Gower, 1971), which can handle both categorical and numerical variables. The legitimate values of this metric range from 0 to 1, where 0 signifies perfect similarity and 1 denotes maximum dissimilarity. Initially, the partial similarity (ps) at each dimension f in the dataset is computed, for each pair of energy efficiency projects. For categorical dimensions, a ps value of 1 is assigned to identical values and 0 to disparate ones. In numerical dimensions, the partial similarity for each dimension f is calculated using a Manhattan distance-based approach, normalized by the range of values recorded in that dimension. Subsequently, the partial similarities across all dimensions are aggregated to calculate the overall similarity between projects. Finally, this cumulative similarity is subtracted from 1 to estimate the total dissimilarity (GD; Eq. (2)).

$$GD = 1 - \frac{1}{m}\sum_{f=1}^{m}\left(1 - \frac{|x_{if} - x_{jf}|}{R_f}\right)^{(f)} \tag{2}$$

where x_{if} and x_{jf} represent the values of energy efficiency projects i and j, respectively, at the numeric dimension f, m denotes the number of dimensions in the dataset, and R_f represents the numeric range in dimension f.

Upon quantifying the dissimilarities between the energy efficiency projects in the dataset, the analysis proceeds to their clustering. For this purpose, the analysis employs the Partitioning Around Medoid (PAM) method (Botyarov and Miller, 2022). This choice is driven by the versatility of this method in handling a custom dissimilarity matrix of dimensions $p \times p$, such as the one generated in the previous methodological step. The underlying rationale of this method hinges on the aspects of medoids, which are actual data points in the dataset, unlike centroids derived from data aggregation. This attribute makes PAM's results more robust to outliers. However, this method comes with certain limitations, which are mainly related to the requirement for predefining the number of medoids and its inherent high computational load. The latter stems from the fact that the method explores all possible combinations across data points for a given number of medoids, assigning each to the closest medoid. Subsequently, the method opts for the combination of data points that minimizes the total distance-based cost (Eq. (3)), optimizing over all potential medoids M.

$$f(M) = -\sum_{j}^{K} d_1(x_j, M) \tag{3}$$

where d represents the distance of the data point x_j from the medoid M of cluster j, and K denotes the total number of clusters.

To determine the optimal number of clusters for the PAM method, the analysis harnesses the Silhouette index. This index is selected for its effectiveness in fitting clusters with observations (Asri et al., 2019). Specifically, it measures how well an element fits in a cluster, by calculating the average dissimilarity a_j of element j to other elements of the same cluster (Eq. (4)). Moreover, it examines what would happen if the element were reallocated to another cluster. This is done by calculating the average dissimilarity b_{jk} of element j to elements I in each different cluster k, except the one to which the element currently belongs (Eq. (5)).

The silhouette index uses the minimum from all the b_{jk} values, along with the calculated a_j value (Eq. (6)). The legitimate values of this index range from −1 to 1, with higher values signifying a more efficient clustering. This index is sequentially calculated for each K value in the [2,10] range. Subsequently, the analysis opts for the number of clusters that yields the highest Silhouette index, and thereby a more efficient clustering. The PAM method is implemented with the "cluster" package in R (Maechler et al., 2021).

$$a_j = \text{avg } d(x_j, x_{j'}), \quad j' \in \{i : l_1(x_i, M) = l_1(x_j, M)\} \tag{4}$$

$$b_{jk} = \text{avg } d(x_j, x_{j'}), \quad j' \in \{i : l_1(x_i, M) = k\} \tag{5}$$

$$S_j(M) = \frac{b_j - a_j}{\max(a_j, b_j)} \tag{6}$$

where $b_j = \min_k b_{jk}$, and $\max(a_j, b_j)$ is the maximum value between a_j and b_j.

2.2.3. Numerical data clustering

In a second stage, the clustering problem is treated with an alternative approach. Specifically, the categorical variables of the dataset are transformed into numerical. This setting in turns allows for applying the K-Means method, which is one of the most efficient in clustering purely numerical datasets (Ping et al., 2024). This transformation is achieved via one-hot encoding approach (Hastie et al., 2009), where each categorical feature is represented by dummy binary variables. In this procedure, each categorical feature's distinct value is mapped with a separate dummy binary variable: 1 is assigned to projects that exhibit this particular value at the examined feature and 0 is allocated to the rest of projects. These dummy variables are then incorporated into the dataset.

The K-means method begins by normalizing all originally numeric variables in the dataset to mitigate the impact of varying scales. K-means requires a predefined number of clusters (K) to classify data. Initially, it selects randomly k objects from the dataset to serve as the preliminary centroids of the clusters. Subsequently, each data point is assigned to its closest cluster to minimize the Sum of Squared Error (SSE). The error for each point is defined as its distance to the nearest cluster center (Eq. (7)) (Hartigan and Wong, 1979).

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{7}$$

where x_i denotes a specific data point in cluster C_k and μ_k represents the centroid of the cluster.

In turn, the cluster centroids are recomputed by designating the mean value of the points constituting each cluster as the new centroid. This iterative process continues until the centroids stabilize (i.e., showing no further convergence), or the predefined maximum number of iterations is reached. To determine the optimal number of clusters, the analysis utilizes in this case the Elbow method (Mehedi Hassan et al., 2022). In the context of this method, the K-means technique is iteratively applied for varying number of clusters, calculating the total WCSS at each iteration (refer to *see* Eq. (7)). In the resultant curve, which maps the total WCSS values against their corresponding cluster counts, the optimal cluster count is observed at the "elbow" point. This point signifies the cluster count beyond which further increasing the number of clusters does not result in a substantial decrease in the WCSS value, despite the added complexity associated with the increasing cluster count. The K-means method is implemented using the "stats" package in R (R Core Team, 2013), setting 10 distinct initial configurations and a cap of 10 iterations.

2.2.4. Evaluation of clustering efficiency

Upon clustering the analyzed energy efficiency projects using the two above-described methods, the study proceeds to assess the robustness of each method's clustering results. In this effort, each method's results are visualized in three dimensions (3D) by applying the t-distributed stochastic neighborhood embedding (t-SNE) dimension reduction technique (Khan et al., 2023). This technique enables the mapping of a high–dimensional space to a low–dimensional one without losing critical information about the connectedness between the elements of different clusters, as well as within the same cluster. It is more adept at preserving local relationships within complex data during dimension reduction compared to linear reduction techniques. Consequently, it allows for visualizing multi-feature clusters in significantly smaller dimensions. In these visualizations, the focus is on examining the compactness of the identified clusters and the extent of inter-cluster overlaps. Effective clustering visualized through t-SNE will comprise tightly grouped clusters, with points within a cluster being as close as possible to each other and as distant as possible from points in other clusters. The t-SNE method is applied using the "Rtsne" package in R (van der Maaten and Hinton, 2008; van der Maaten, 2014).

2.3. Assessment of feature importance in the classification model

2.3.1. Methodology

After clustering the energy efficiency projects using the above-described methods, the analysis seeks to identify the key factors that determine a project's classification into a specific cluster. Subsequently, the study aims to discern the factors associated with the successful implementation of energy efficiency projects. For this purpose, the analysis employs the Random Forest method for classification (Metzig et al., 2020), which has been proven to be among the most accurate methods for classification tasks (Jiang et al., 2023).

Random Forest fits the scopes of the analysis since, as an ensemble method, it derives its outcomes by aggregating the results of multiple decision trees, thereby minimizing the risk of overfitting (see Appendix A for the mathematical formulas). This is achieved through bagging, a process involving the random selection of subsets of the dataset (both rows and columns) with replacement. Replacement means that any given data point can feature in several subsets. A decision tree (Koutsandreas, 2023) is then constructed for each subset, and the final predictions hinge on the majority voting across these decision trees. Fundamentally, Random Forest is more robust to outliers compared to other ensemble model alternatives (e.g., Gradient Boosting) as it constructs decision trees in parallel rather than sequentially. This approach eliminates dependencies between decision trees, making it more robust to outliers. This feature is particularly important in this analysis as it deals with a heterogeneous dataset with multiple features and large discrepancies between projects (*see* Fig. 1). Additionally, Random Forest can handle non-linear relationships between independent and dependent variables more effectively compared to linear-based classification models (e.g., logistic regression). It should be noted that the extent of the utilized dataset would not suffice for training neural networks, thereby compromising their efficiency.

Furthermore, Random Forest facilitates the evaluation of feature importance, which is the main rationale behind conducting this methodological stage. This makes it preferable compared to other classification model alternatives that lack methods for direct feature importance evaluation (e.g., Support Vector Machines). Particularly, feature importance is evaluated by calculating the average decrease in the Gini Impurity index across the constructed decision trees for each feature. This index measures the impurity of a decision tree node when using a particular feature to split it. Therefore, the extent to which a feature reduces impurity across decision trees indicates its influence on the classification model.

A Random Forest model is trained to accurately project an energy efficiency project's cluster based on its characteristics. The mapping of projects to clusters is derived from the results of the previous stage of the analysis, based on the results of the PAM and K-means methods. Following the model's training, the relative importance of the dataset's features is calculated to evaluate their significance in the classification model. This methodological stage is implemented using the "randomForest" package in R (Liaw and Wiener, 2002).

2.3.2. Sensitivity analysis of results across classification models

As mentioned above, the features of the Random Forest model make it suitable for the scopes of this analysis, particularly for discerning the influential factors of the classification scheme. The factors behind the suitability of this model compared to its main alternatives, can be summarized to the following points (see above for details): (i) Random Forest can handle outliers better than Gradient Boosting; (ii) Random

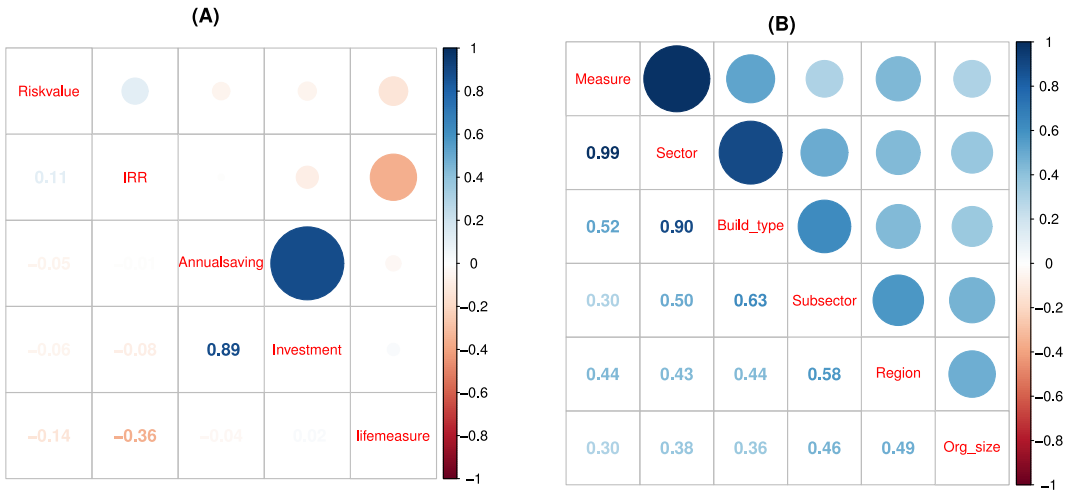

Fig. 3. Visual representation of correlations between the dataset's features. Panel A illustrates the correlations between the dataset's numerical features in terms of Pearson's coefficients. Panel B visualizes the correlations between the dataset's categorical variables in terms of Cramer's V values. The strength of correlations is indicated by circle color and size above the main diagonal and numbers below it.

Forest is more suited to deal with non-linear data compared to Logistic Regression; Random Forest includes a direct feature importance evaluation method, unlike Support Vector Machines; Random Forest is apt for smaller datasets, whereas neural networks necessitate substantial data volumes for effective training.

To verify the principal superiority of Random Forest over other classification models, under the conditions of this study, the analysis performs a sensitivity analysis of results to the chosen classification model. This analysis considers the most commonly used classification models, including Logistic Regression, Support Vector Machines, and Gradient Boosting Machine, alongside Random Forest. The detailed mathematical formulas for these models are presented in Appendix A to avoid distracting the main focus of the paper and due to space limitations. Subsequently, the analysis computes the rank errors in feature importance each classification model produces from the perspective of the other models, as defined in Eq. (8). In this process, the errors produced by each classification model for each feature across the other models are summed, as well as the errors each model produces across all features. By doing so, the analysis aims to identify the "least wrong" classification model from the perspective of the competing models.

$$\left(\sum_{c=1}^{k}\sum_{j=1}^{n}\left|x_m - c_j\right|\right)_s \tag{8}$$

where S includes the classification models the analysis applies ($S = \{\mathrm{RF}, \mathrm{LR}, \mathrm{SVM}, \mathrm{GBM}\}$), k denotes the features of the dataset (see Table 1 for details), and n accounts for the set of examined classification models excluding the model m for which the errors are being calculated each time ($n = \{S\} \setminus \{m\}$).

3. Results analysis

3.1. Pre-clustering results

This section presents the results of the analysis performed to lay the groundwork for an efficient clustering. This includes assessing feature correlations and cluster efficiency across various cluster counts. Feature correlations analysis is performed to identify features that may be excluded from the cluster analysis in light of eliminating data distortions and biases. Fig. 3 illustrates the correlations between the numerical features of the dataset in terms of Pearson's correlation coefficients (Panel A) and the categorical features of the dataset in terms of Cramer's V values (Panel B). The strength of correlations is indicated by the circle size and color above the main diagonal and numbers below it.

Regarding the numerical variables of the dataset, as shown, "Annual Savings" variable is strongly correlated with the one denoting "Initial Investment". Therefore, the "Annual Savings" variable is opted for exclusion from the dataset between the two before proceeding with the clustering of the data. This decision is driven by the fact that the "Initial Investment" variable contains important information about the scale of the project, not directly captured by another variable of the dataset (unlike IRR for savings, for example).

Regarding categorical features, there is a strong correlation between the variable describing each project's sector and the variables describing the type of intervention and associated building type; while the correlation between the latter two variables is rather moderate. Hence, the "project sector" variable is excluded before proceeding with data clustering. It is important to note that the project sector is also captured by the "Building type" variable. For this reason, "Building type" is renamed as "Sector–Building type".

Going forward, Fig. 4 illustrates the results about the Silhouette index and WCSS values, obtained from clustering the analyzed energy efficiency projects using the PAM and K-means methods, respectively, across various cluster counts. This analysis helps determine the optimal number of clusters per each method utilized. Particularly, Panel A displays the silhouette index values obtained from the application of the PAM method within the range of two to ten clusters. It should be noted that higher silhouette index values indicate a more efficient clustering. Conversely, Panel B presents the WCSS values obtained by clustering the analyzed energy efficiency projects with the K-means method, incrementally considering a number of clusters from one to ten.

As shown, the highest Silhouette index values for the PAM method arise when the number of clusters is at the two extremes of the examined range, namely two or ten (Panel A; Fig. 4). On the other hand, the 'elbow' point in the WCSS graph for the K-means method is observed around three clusters (Panel B; Fig. 4). Applying the K-means method for more than three clusters would not result in a significant decline in the within-cluster variance, despite the additional complexity from the increased number of clusters. To maintain a consistency between the two methods and perform a more nuanced clustering, both the PAM and K-means methods are applied for three clusters. This decision is further supported by the fact that the decrease in the Silhouette index when considering three, instead of two, clusters is relatively minor. This approach allows for a direct comparison of the results obtained from both methods.

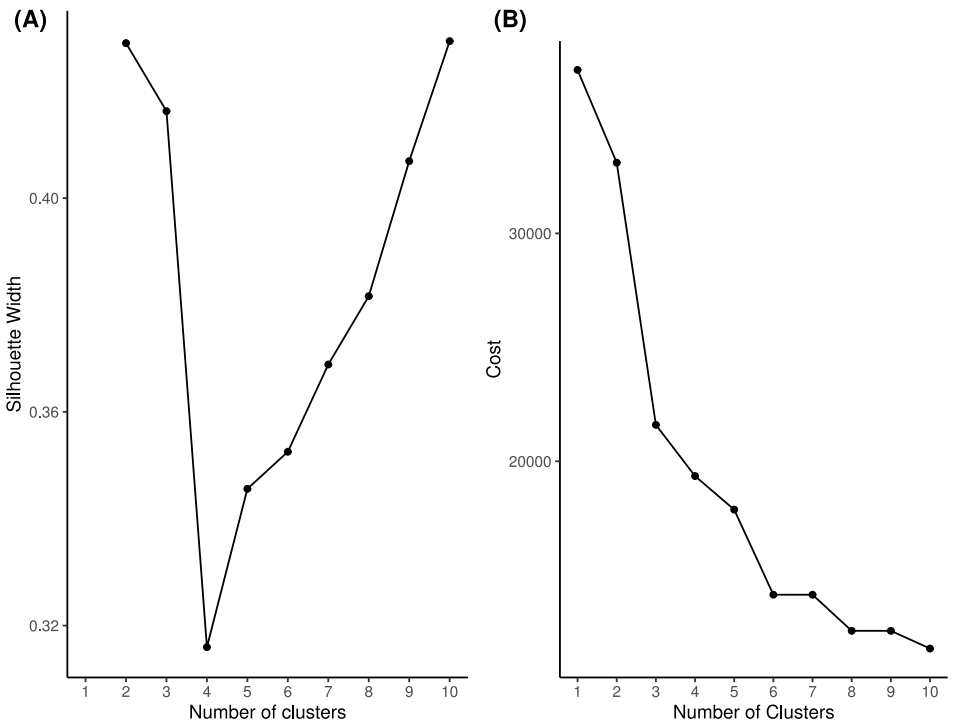

Fig. 4. Evaluation of the clustering efficiency across various cluster counts. Panel A displays the Silhouette index values derived from clustering the analyzed energy efficiency projects with the Partitioning Around Medoids (PAM) method in the range of two to ten clusters. Panel B illustrates the total within-cluster sum of squares (WCSS) values obtained by clustering the analyzed energy efficiency projects with the K-means method in the range of one to ten clusters.

3.2. Clustering results

This section presents the main results obtained by implementing the two adopted clustering approaches with three clusters, following the exclusion of correlated features from the dataset. The clustering results for each method are visualized in three dimensions in Fig. 5, with Panel A representing the PAM method and Panel B accounting for the K-means method. These visualizations, created using the t-SNE dimension reduction technique, serve as indicators of clustering efficacy. Typical characteristics of an effective clustering include the proximity between points belonging to the same cluster and the absence of intra-cluster overlaps.

Although hinging on a substantial dimension reduction, the visualizations in Fig. 5 reveal that the K-means method achieves more compact clustering with fewer intra-cluster overlaps than the PAM approach. This outcome is influenced by the fact that the PAM method was applied for a near-optimal number of clusters, as determined by the Silhouette index, to maintain consistency between the two employed methods. It is important to note that each clustering approach fundamentally reflects a different methodological way of normalizing nominal data. Therefore, they do not have a particular meaning regarding the characteristics of energy efficiency investments and how these are treated. In the subsequent figures, a more nuanced evaluation of the differences between these two approaches is provided.

Delving further into the results, Fig. 6 provides an indication of the characteristics of the energy efficiency investments that form the clusters. In particular, it illustrates the median values within each cluster for the risk of failure (Panel A), project IRR (Panel B), initial investment (Panel C), and life measure (Panel D), for both the PAM (solid lines) and K-means (dotted lines) methods. In the same context, Fig. 7 displays the distributions at each cluster across the assumed regions (Panel A) and energy efficiency measures included in the dataset (Panel B), for both the PAM and K-means methods.

Fig. 6 reveals minimal differences between the results of the two employed clustering methods, whereas, in some features, such as risk, their results become nearly identical. As for differences between clusters in each method, they are mainly observed in risk and IRR, while in the remaining features, there are differences only between Cluster 1 and the other clusters, with Clusters 2 and 3 displaying similar attributes. Digging deeper, Cluster 1 investments are characterized by a rather moderate risk coupled with low profitability. Interestingly though, these investments require the highest initial investment and present the longest payback period among the clusters. The latter characteristic implies a delayed return on investment, making these investments less lucrative. Putting things into perspective, the attributes of Cluster 1 investments are mainly driven by their focus on "Building Fabric Measures", which display lengthy lifetimes and low profitability. Additionally, the prevalence of this investment type in Eastern European countries, whose economies lag behind those of the other examined regions, may also contribute to these attributes.

On the other hand, Cluster 2 investments exhibit a quite significant profitability along with minimal risk. These are smaller-scale investments in terms of initial capital and have a considerably shorter payback period than Cluster 1. Concerning the intervention and regional distribution of this investment class, it is mainly made of "HVAC Plant" and "Lighting" energy efficiency measures, which are mainly found in Western European countries, with a minor presence in Northern Europe. As for Cluster 3 investments, they display the highest profitability but also bear the greatest risk among the examined clusters. These investments are chiefly executed in Western Europe and, to a lesser extent, in the USA. They also require a rather similar initial investment to Cluster 2 investments, which is substantially lower than that of Cluster 1. A distinct feature of Cluster 3 is its diverse mix of energy efficiency measures, which include, inter alia, "Heating", "Cooling", "Power systems & Motors", "Compressed Air", and "Waste Heat".

It is noteworthy that investments in Cluster 2, and particularly in Cluster 3, primarily belong to the Industry sector, whereas Cluster 1 investments are predominantly related to the Building sector, especially household buildings. This distinction can be seen in panel B of Fig. 8, which illustrates the distribution of projects across building types for

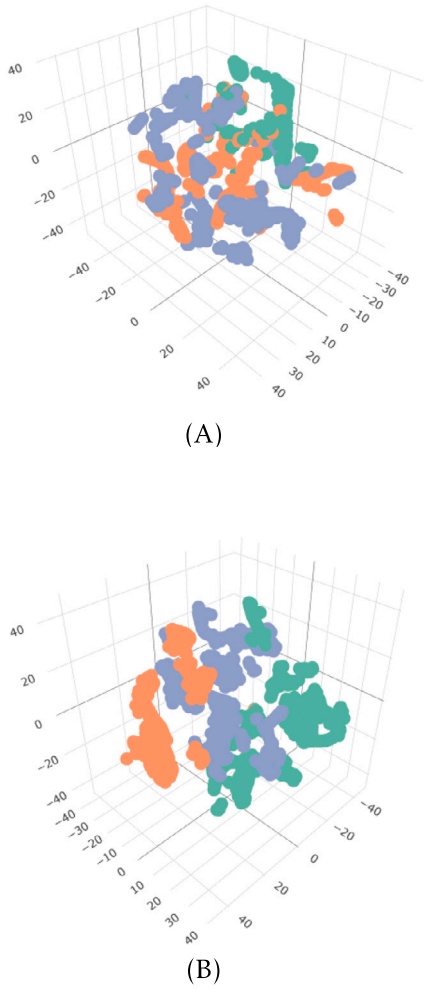

Fig. 5. Three-dimensional visualization of performed clustering of analyzed projects using the t-distributed stochastic neighborhood embedding (t-SNE) dimension reduction technique, for both the PAM (Panel A) and K-means (Panel B) methods. Points belonging to the same cluster are marked with identical colors.

each cluster, in both the PAM and K-means clustering cases. This trend explains much about the profitability of investments in Clusters 2 and 3. This is because industry-focused interventions yield higher returns for investors, notably due to the more advanced involved technologies (Kleanthis et al., 2022).

Notably, two of the three identified clusters (Clusters 1 and 2) predominantly consist of certain types of interventions that define their characteristics. These projects exhibit distinct connections with other features, such as risk or building type, which ultimately play a crucial role in a project's categorization. For instance, such features significantly impact whether projects involving building fabric measures are classified as no-viable investments in Cluster 1 or as viable investments in Cluster 2. Therefore, the types of interventions comprise the first and foremost classification mechanism between clusters. Following this, the interplay of these energy efficiency measures with other project attributes like region, building type, etc., affects the total risk and profitability of the investments and ultimately their final cluster classification.

On the other hand, panel A of Fig. 8 illustrates the cluster membership of investments, expressed as the percentage of projects belonging to each cluster relative to the total number of projects. This visualization fortifies the observation that the PAM and K-means clustering results are largely similar. A minor difference is that the PAM method classifies more investments into the lower-risk, higher-profitability Cluster 2: these additional investments, identified as the most profitable in Clusters 1 and 3 as per the K-means method, are assigned to Cluster 2 by the PAM method. This reassignment results in slightly higher profitability for PAM's Cluster 2 and, correspondingly, lower profitability for PAM's Clusters 1 and 3, compared to the K-means method. However, the central risk values remain almost similar between the two clustering methods (Fig. 6).

Therefore, the analysis suggests that energy efficiency investments can be, by and large, classified into three categories. Among these, two can be considered go-investments by energy efficiency services market stakeholders. Investors should choose between these two categories according to their preferences, such as risk tolerance, profitability goals, and available investment capital. For instance, more conservative investors might prefer Cluster 2 investments due to their high profitability and low risk, henceforth referred to as "safe profitability". On the other hand, more risk-taking investors may opt for Cluster 3 investments, in light of increasing their portfolio profits at the expense of undertaking a higher risk. These investments are labeled as "high stakes". It is noteworthy that real-world investors usually diversify their portfolios with a mix of investment categories, striking a balance between risk and profitability according to their preferences. Finally, Cluster 1 investments, characterized by low returns over a long horizon in tandem with a notable risk of failure, might be considered no-go investments: putting money on these investments entails opportunity costs, as the same capital could be directed to other more profitable investments with similar or lower risk levels. The analysis colloquially terms these investments as "junk".

Concerning regional classification, it is noteworthy that Western European projects dominate the dataset (see Subsection 2.1), something that in turn influences clusters' regional distribution (Fig. 7) and characteristics. Specifically, Clusters 2 and 3, mainly composed of projects implemented in Western Europe and the USA—regions offering more favorable economic conditions—exhibit the highest profitability. The extent to which this trend is influenced also by other features of the projects at hand (e.g., intervention measures) will be elucidated by the results of the Random Forest model. It should be noted that various factors inherent to regional conditions can significantly affect the profitability of energy efficiency projects implemented within the same territory (Gillingham and Palmer, 2014; Koutsandreas et al., 2022). First, energy prices, dictated by regional markets, are a fundamental factor for such projects as they directly influence returns, along with estimated energy savings. In advanced economies, energy prices may be higher due to increased demand from more energy-intensive lifestyles, or additional costs such as environmental levies. Furthermore, these economies often offer specific incentives for such projects, like tax rebates, which can reduce the initial investment cost. Additionally, the higher availability of skilled professionals and access to more advanced technology can lead to increased energy savings and subsequently higher returns.

Regarding this model, Fig. 9 illustrates the average decrease in the Gini index across the decision trees of the trained Random Forest model for each feature used for splitting. This index denotes the impurity in the decision tree when a particular feature is used for splitting. Therefore, a higher decrease in the Gini index is associated with a greater influence of the feature in question on the classification model. Results are visualized for both clustering cases (i.e., PAM- and K-means-based), including the average values between these two methods. The length of the lines in Fig. 9 reflects the sensitivity of the results to the clustering approach of choice.

As demonstrated, both clustering methods concur on the least influential factors in the classification model. However, there are some discrepancies regarding the most influential features and those with intermediate influence, depending on the chosen clustering approach, which may affect the relative ranking of features. Nevertheless, the general picture of the key factors influencing the classification scheme remains consistent across both clustering approaches. Specifically, the

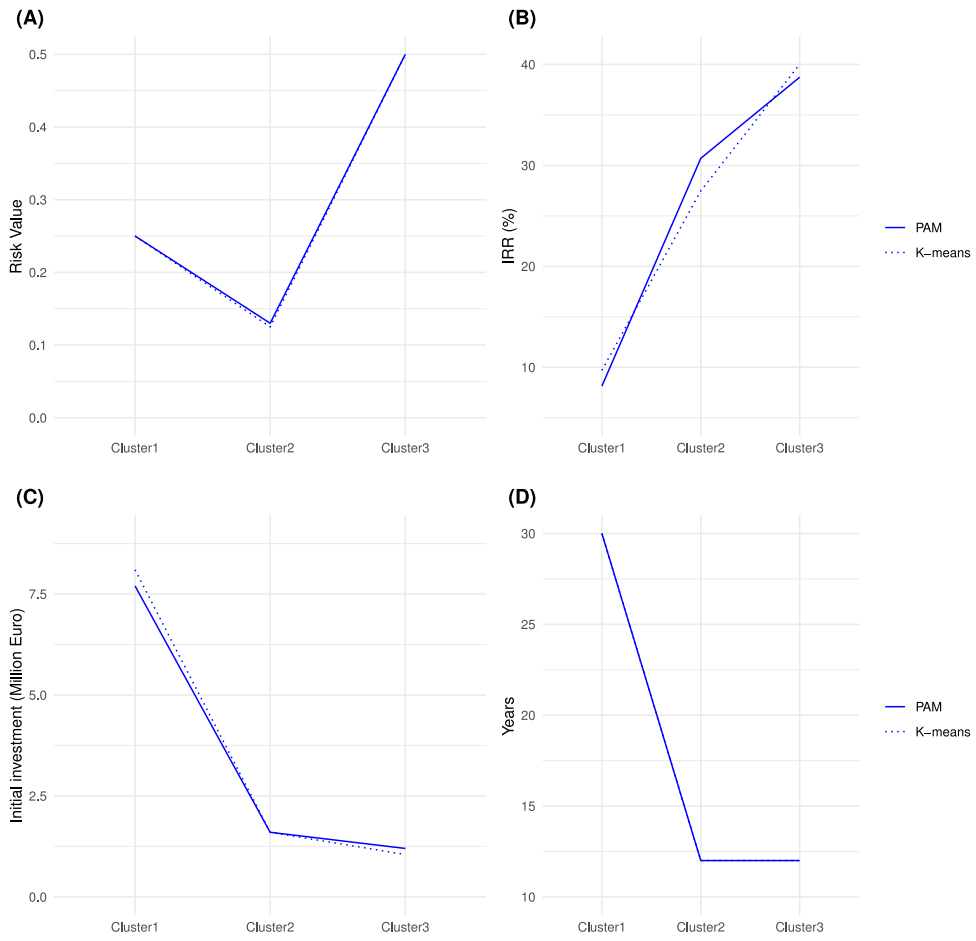

Fig. 6. Median values of clustered energy efficiency projects in terms of risk of failure (Panel A), project IRR (Panel B), initial investment (Panel C), and life measure (Panel D) across the PAM (solid lines) and K-means (dotted lines) methods.

analysis reveals that the risk of failure, energy efficiency intervention, and project sector are the most impactful factors in the trained Random Forest-based model, although their influence varies with the selected clustering method. In this context, while the PAM–derived results identify the project sector as more influential factor than energy efficiency intervention, the K-means-derived ones indicate the contrary.

Regarding features associated with project profitability, such as initial investment and IRR, they generally rank lower in terms of influence, with a high level of agreement between the two employed clustering methods. An exception is the project life measure, which has a rather moderate impact that significantly depends on the chosen clustering method. A similar pattern is observed for the project region of implementation, with it being more influential when clustering is performed with the PAM—versus the K-means—method. Furthermore, the organization size and project subsector do not emerge as highly influential factors in the classification scheme.

However, it should be noted that features not critical for classification, such as region of implementation, can still overwhelmingly affect the characteristics of the classes. This is evident in Cluster 1, where most projects are located in Eastern Europe, especially under the K-means case. This suggests that a cluster can primarily reflect one specific measure with distinct links to other characteristics (e.g., region, risk). Consequently, the features that most effectively describe these connections arise as the most influential in the classification model. It should be noted that the examined indicators bias the strongest connections identified. For instance, the prevalence of economic factors in the dataset results in their strong role in the formulated classification schemes.

To sum up, the analysis indicates that, when examining energy efficiency projects, capital providers and research community should consider not only traditional profitability indicators but also those reflecting projects' risk of failure and technical characteristics (e.g., building type). That said, while traditional indicators, such as IRR and region of implementation, effectively categorize different investments, factors like risk of failure can be crucial in determining viable investments, especially for those lying on the verge of being viable according to profitability indicators.

3.3. On the effect of classification model's selection on feature importance

This subsection presents the key results about the sensitivity of feature importance to the selected classification model. Specifically, Fig. 10 illustrates the divergence in feature importance ranks across different classification models. Each model's ranks are highlighted with a different marker.

As shown, the results feature a high sensitivity to the chosen classification model. This is not something unexpected given the different underlying rationales of these models (see the discussion above). However, the various classification models agree on the most and least influential features for classification, especially when classification models are trained with the K-means clustering results. The only exception is Logistic Regression model, whose results significantly diverge from the others — this verifies the principal inability of this model to effectively deal with non-linear data patterns. Random Forest results are very close to those of Gradient Boosting — something expected, as the models have similar underlying rationales, differing primarily in how

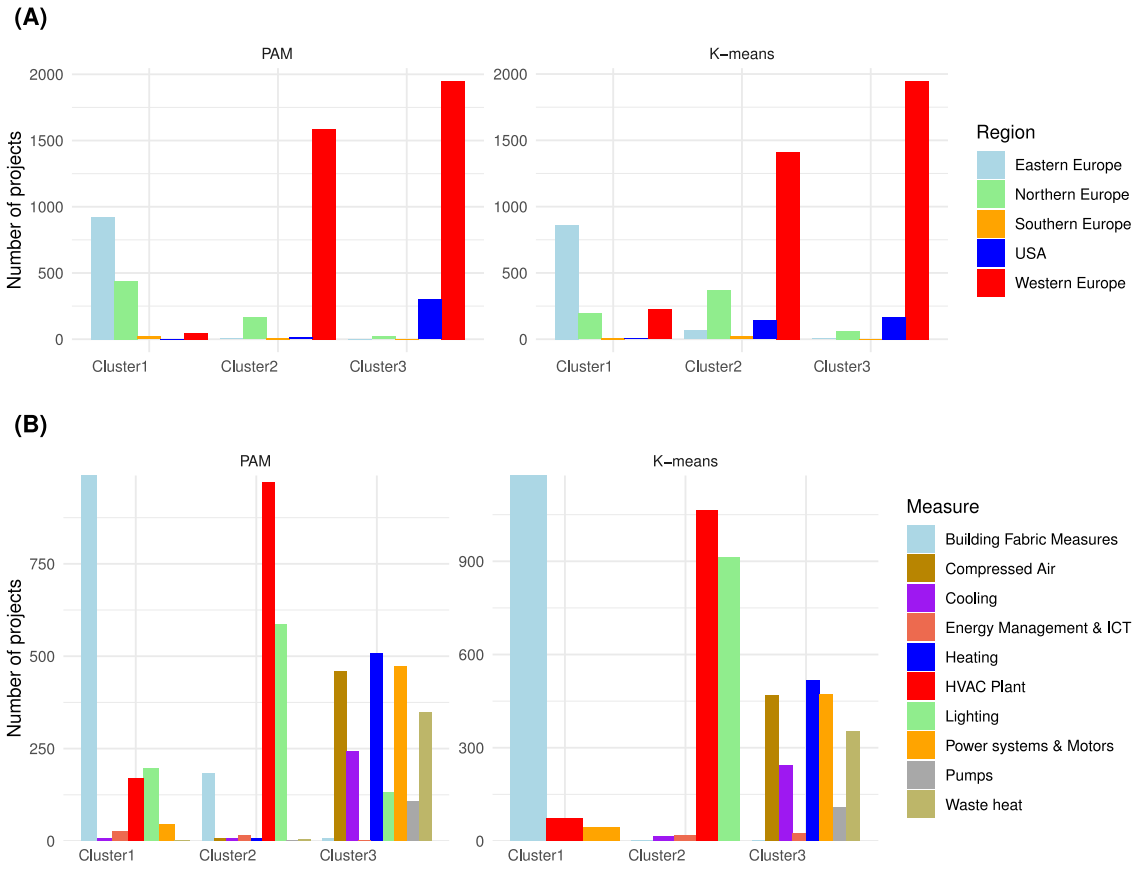

Fig. 7. Distribution of clustered energy efficiency projects across regions (Panel A) and intervention measures (Panel B), for both the PAM and K-means methods.

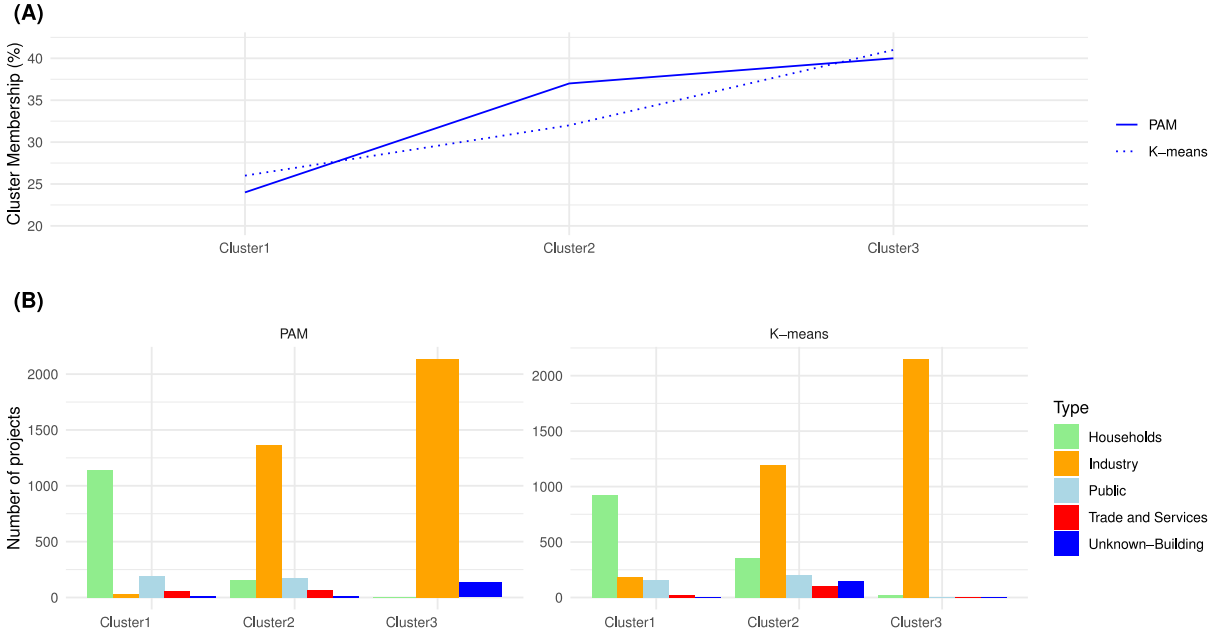

Fig. 8. Characteristics and structure of energy efficiency investment clusters. Panel A illustrates the investment cluster membership as a percentage of projects in each cluster relative to the total number of projects, for both the PAM (solid line) and K-means (dotted line) methods. Panel B represents the distribution of clustered energy efficiency projects across building types, for both the PAM and K-means methods. The "Unknown–Building" value denotes projects belonging to the Building sector but with an unspecified associated building type.

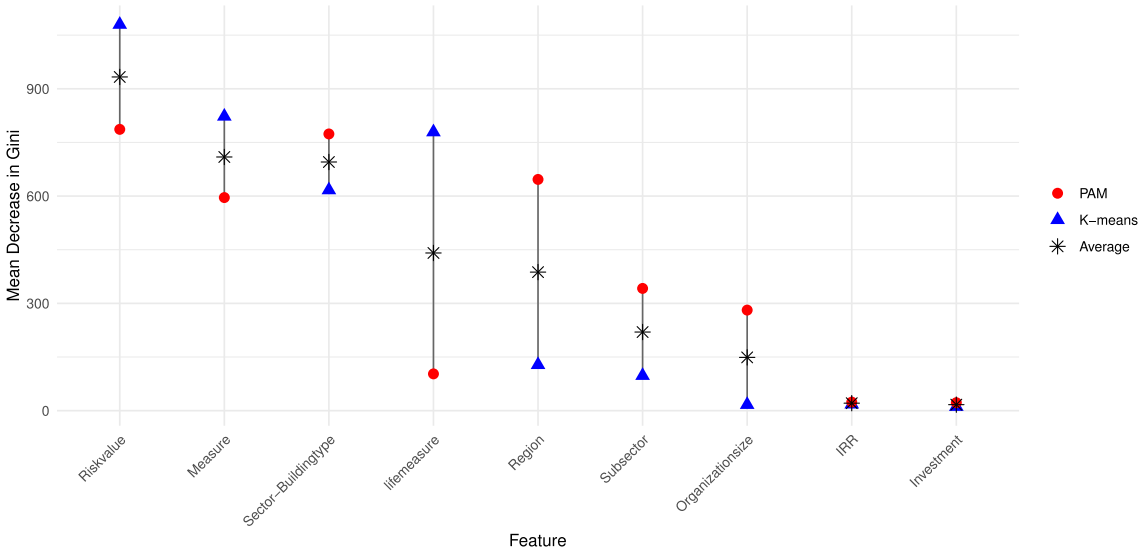

Fig. 9. Visual illustration of the average decrease in the Gini index in the decision trees of the trained Random Forest model across the features used for splitting, for both the PAM- and K-means-derived results including their average values. A larger decrease signifies a stronger influence of the features on the classification model.

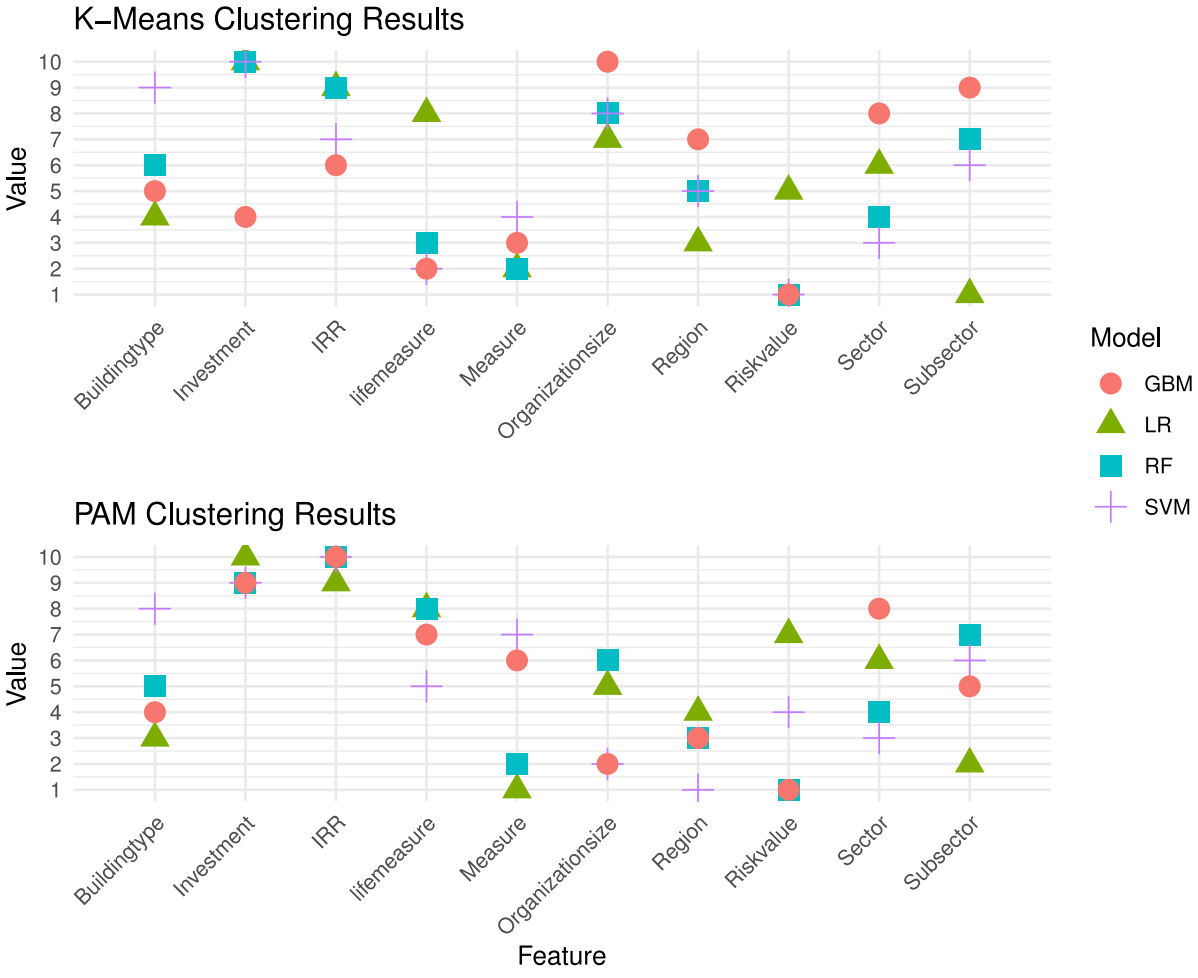

Fig. 10. Variability of feature importance ranks across different classification models, including Random Forest (RF), Logistic Regression (LR), Gradient Boosting Machine (GBM), and Support Vector Machine (SVM). Each method's results are marked with a different symbol.

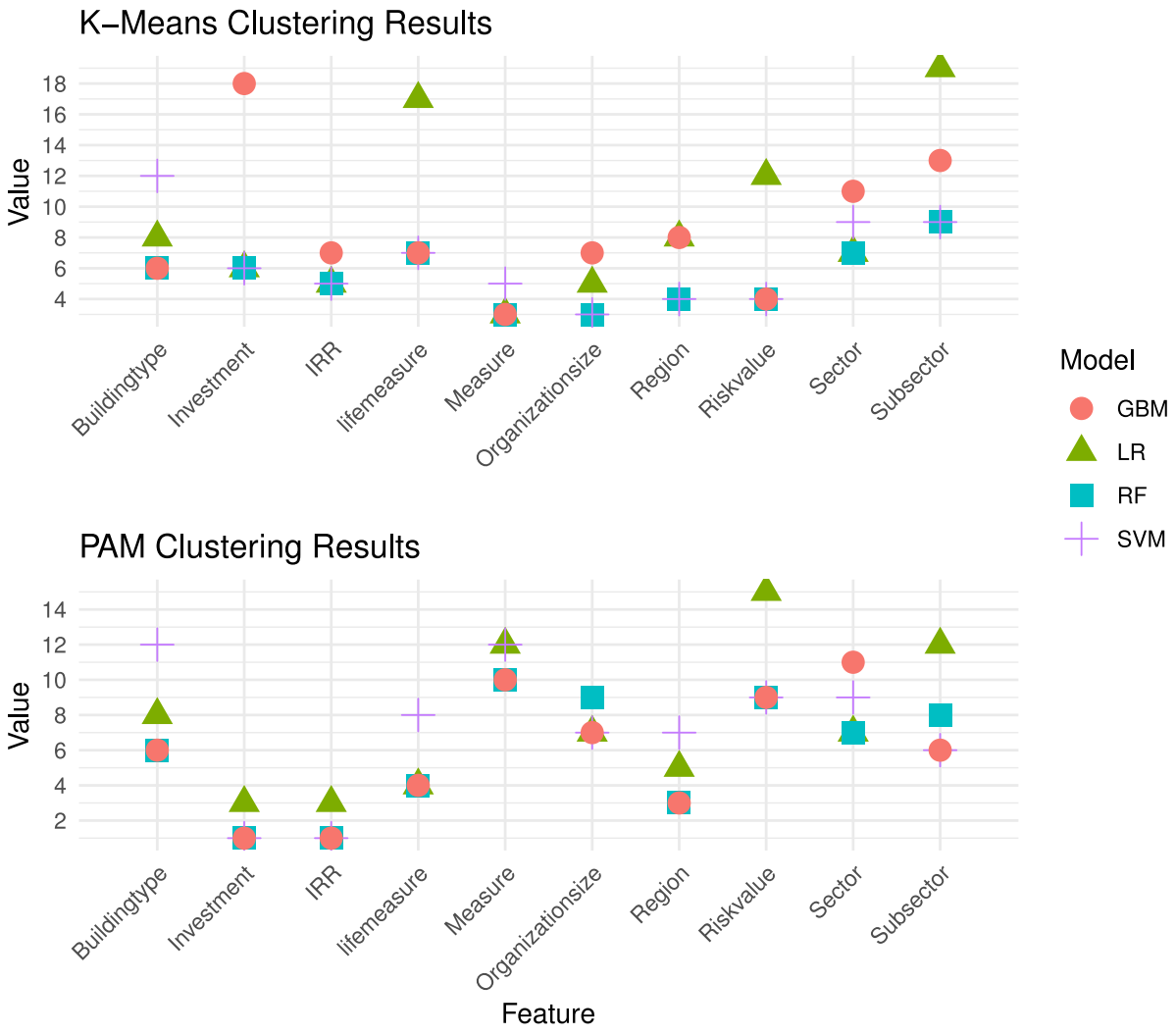

Fig. 11. Errors in the form of absolute differences between the ranks produced by each classification model and the ranks produced by the other classification models (Eq. (8)).

decision trees are constructed, namely in parallel for Random Forest or sequentially for Gradient Boosting.

As mentioned above, Random Forest is, in principle, the most suitable model for evaluating feature importance in this study. This is further supported by the results presented in Fig. 11, which illustrates the rank errors each classification model produces from the perspective of the other classification models. As this figure reveals, Random Forest produces ranks that are, on average, the "least wrong" from the perspective of other models — this is the case for all individual factors with K-Means and for all but two with PAM. This outcome underscores the robustness and reliability of the results obtained using this model. Additionally, the results emphasize the influence of risk of failure in the classification scheme, as it is consistently recognized as the most influential factor across models with varying rationales.

Appendix B provides additional results from the sensitivity analysis concerning classification models. Specifically, Fig. B.1 presents the normalized feature importance across the different classification models. The normalized performance is calculated by scaling the importance value that each model assigns to a given feature by the sum of the values this model generates across all features — so that the sum of normalized performance values across features for each model equals 1. These results further highlight the high sensitivity of feature importance to the chosen model, which becomes more pronounced when focusing on the normalized importance of features across each method, rather than solely on the ranks produced by them. This is related to the different rationales and scales of the results generated by different methods.

4. Conclusions

Bridging the transparency gap in energy efficiency investments is crucial to enhancing their attractiveness, thereby increasing project implementation. This paper aims to bridge this gap by identifying families of energy efficiency investments and the determinant factors of cluster membership, using machine learning algorithms. The analysis harnesses a broad spectrum of indicators about successfully implemented projects in Europe and the USA, while utilizing two clustering approaches, namely PAM and K-means, for three clusters. The influence of features in the classification scheme is assessed using a Random Forest model, in turn calculating the average decrease in the Gini index across features.

The comparison of results from the two employed clustering approaches reveals only minor differences, highlighting the robustness of the analysis. This applies especially to the characteristics of the clusters and, less so, to the influential factors of the classification scheme, of which relative ranking can be overturned according to the chosen clustering approach. Analysis results suggest that energy efficiency investments can be largely classified into three key classes. The first, colloquially termed "junk", involves low profitability projects (IRR~10%) with moderate risk, high initial investments, and extended horizons. The second category, labeled as "safe profitability", is mainly

composed of high–profitability investments (IRR∼30%) coupled with minimal risk, and short horizons. The final category, entitled as"high stakes", involves very high-profitability investments (IRR∼40%) along with short horizons, though accompanied by a considerable risk.

The latter two investment categories can be seen as viable options by profit-focused investors, who should choose between the two based on their preferences, such as their attitude against risk, preferred investment horizon, and available investment capital. In the proposed classification scheme, the primary influential factors include the risk of failure, energy efficiency intervention, and building type (which also indicates the investment sector), followed by life measure and region of implementation. While purely profitability-related indicators, such as IRR and initial investment, emerge as insignificant in the classification scheme, they still have a considerable impact on the characteristics of the classes. The sensitivity analysis revealed high discrepancies across different classification models and supported the robustness of Random Forest in assessing feature importance. The analysis suggests that capital providers should shift their focus from solely evaluating such investments from a profitability perspective to also considering the associated uncertainties and inherent technical aspects, such as the involved building type. This is crucial as these often-neglected characteristics can be the deciding factors in a project's viability, especially for those with poor or moderate profitability. Furthermore, the analysis outcomes indicate that policymakers should formulate support mechanisms and policy packages for similar groups of energy efficiency investments rather than at the intervention and regional levels. This approach can help reduce the heterogeneity of energy efficiency policies, thereby stimulating the upscaling of such investments. Moreover, this framework can target the investments that most need support, either because they present high risk, exhibit low profitability, or require high initial investment.

This analysis comes with certain caveats. First, the characteristics of clusters are determined by the specific dimensions used to examine energy efficiency projects. For example, the prevalence of economic indicators in the dataset makes them inherently significant for project classification. Therefore, potential ways forward for the analysis could involve investigating additional features associated with the implementation of energy efficiency projects. Additionally, a finer risk evaluation of energy efficiency projects, such as by incorporating granular information about projects (e.g., experience of involved employees), could lead to a more nuanced clustering. This approach could also be enhanced by incorporating more advanced scientific methods, such as probabilistic risk analysis or scenario analysis.

Furthermore, employing additional classification models could help inspect the sensitivity of the results to the classification model of choice. Finally, both clustering approaches are fed with the optimal number of clusters. Although systematic approaches are used to determine this number, their results are sensitive to the initial parameter settings. Additionally, the same cluster count was considered for both methods to maintain consistency and results comparability, although they yielded a slightly different optimal cluster count. Future research should explore results sensitivity when using different, possibly near-optimal cluster numbers.

CRediT authorship contribution statement

Diamantis Koutsandreas: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ilkka Keppo:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

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

Data availability

The data that has been used is confidential.

Appendix A. Mathematical formulas for classification models

Random Forest (RF)

A Random Forest model is trained to predict an energy efficiency project's cluster based on its characteristics. It derives its outcomes by aggregating the results of multiple decision trees based on the majority vote (Eq. (A.1)).

$$\hat{y} = \text{mode}\left(\{y_i^{(1)}, y_i^{(2)}, \dots, y_i^{(B)}\}\right) \tag{A.1}$$

where $y_i^{(b)}$ is the prediction of the bth tree for the ith sample.

Random Forest assesses feature importance by averaging the reductions in Gini impurity for classification that a feature produces across all trees when it is used for splitting (Eq. (A.2)).

$$\text{Importance}(X_m) = \frac{1}{B}\sum_{b=1}^{B} \Delta\text{Impurity}(X_m, T_b) \tag{A.2}$$

where $\Delta\text{Impurity}(X_m, T_b)$ indicates the impurity decrease by feature X_m in tree T_b.

The Random Forest model is implemented using the `randomForest` package in R. Once the model is trained, feature importance is extracted using the `importance` function from the `randomForest` package.

Logistic Regression (LR)

Logistic Regression is applied to binary classification problems, estimating the probability that an energy efficiency project $\mathbf{x}$ belongs to a specified class as the sigmoid function of a linear combination of the input features (Eq. (A.3)). This method is applied separately for each of the three classes.

$$p(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \tag{A.3}$$

where $\beta_0, \beta_1, \dots, \beta_n$ denote the parameters of the model, and $\mathbf{x} = (x_1, \dots, x_n)$ account for the input features.

The method is implemented in R using the `multinom` function of the `nnet` package. To evaluate feature importance in the classification scheme, the coefficients for each class relative to a baseline class are calculated. These coefficients indicate the change in the log-odds of being in a specific class versus the baseline class for a one-unit change in the predictor. Since there are three classes, two sets of coefficients are computed, which are then averaged. Additionally, for the categorical variables, coefficient are calculated at a value's level, which are again averaged at the category's level to identify the most influential factors in predicting the cluster.

Support Vector Machine (SVM)

Support Vector Machine is trained on the identified classification scheme of energy efficiency projects. Support Vector Machine fundamentally search for the hyperplane that best divides a dataset into two classes.

The decision function for the optimal hyperplane using the kernel function is presented in Eq. (A.4).

$$f(\mathbf{x}) = \text{sgn}(\sum_{i=1}^{m} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b) \tag{A.4}$$

where $y_i \in \{-1, 1\}$ are the class labels, α_i are the Lagrange multipliers for the support vectors $\mathbf{x}_i$, $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function, and b is the bias.

This method is implemented in R using the `e1071` and `caret` packages, with the Support Vector Machine model trained using the `svm` function and tuned via the `train` function for optimized hyperparameters. To evaluate feature importance, a permutation-based approach is used. This approach measures the increase in the prediction error of the model after permuting each feature, providing insight

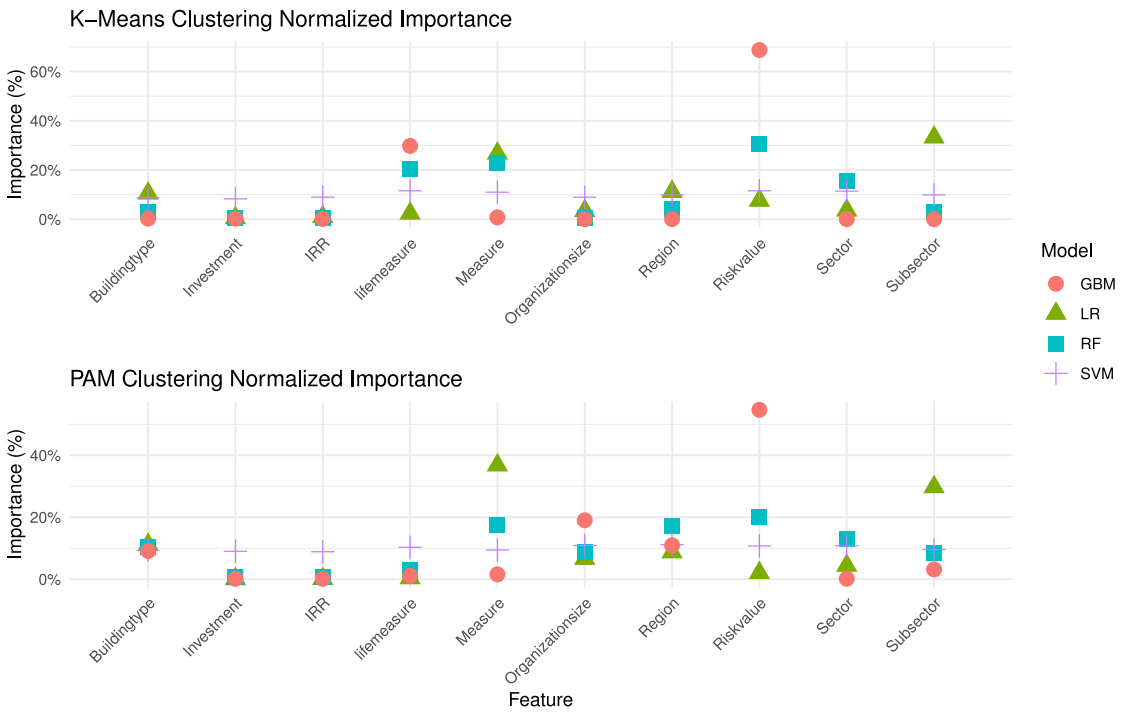

Fig. B.1. Normalized feature importance across the classification models used in the analysis. Different classification models are highlighted with different markers.

into which features most significantly impact model accuracy. Feature permutation involves randomly shuffling the values of a feature across all the observations in the dataset.

Gradient Boosting Machine (GBM)

A Gradient Boosting Machine model is trained to predict the cluster an energy efficiency project belongs to according to its attributes. The model is updated when a new learner is added, as per Eq. (A.5).

$$F_{t+1}(\mathbf{x}) = F_t(\mathbf{x}) + \nu \cdot h_t(\mathbf{x}) \tag{A.5}$$

where $F_t(\mathbf{x})$ is the ensemble model at iteration t, $h_t(\mathbf{x})$ is the weak learner fitted on the negative gradient of the loss function at iteration t, and ν is the learning rate, which scales the contribution of each weak learner.

The Gradient Boosting Machine model is implemented using the `gbm` and `caret` packages in R, where the model is trained and tuned through the `train` function with various parameters specified in a tuning grid to optimize performance.

Once the model is trained, feature importance is extracted using the `varImp` function from the `caret` package. This function computes the importance of each feature in the model as the sum of the improvements in accuracy brought by a feature across all trees in the model where it appears.

Appendix B. Normalized feature importance across classification models

See Fig. B.1.

References

Ahmad, A., Khan, S.S., 2019. Survey of state-of-the-art mixed data clustering algorithms. IEEE Access 7, 31883–31902.

Al Kez, D., Foley, A., Abdul, Z.K., Del Rio, D.F., 2024. Energy poverty prediction in the United Kingdom: A machine learning approach. Energy Policy 184, 113909. http://dx.doi.org/10.1016/j.enpol.2023.113909, URL https://www.sciencedirect.com/science/article/pii/S0301421523004949.

Asri, H., Mousannif, H., Al Moatassime, H., 2019. Reality mining and predictive analytics for building smart applications. J. Big Data 6, 1–25.

Babu, S.C., Gajanan, S.N., 2022. Chapter 4 - effects of technology adoption and gender of household head: the issue, its importance in food security—application of cramer's v and phi coefficient. In: Babu, S.C., Gajanan, S.N. (Eds.), Food Security, Poverty and Nutrition Policy Analysis (Third Edition), third ed. Academic Press, San Diego, pp. 105–133. http://dx.doi.org/10.1016/B978-0-12-820477-1.00019-X, URL https://www.sciencedirect.com/science/article/pii/B978012820477100019X.

Belenguer, A., Pascual, J.A., Navaridas, J., 2023. GöwFed: A novel federated network intrusion detection system. J. Netw. Comput. Appl. 217, 103653. http://dx.doi.org/10.1016/j.jnca.2023.103653, URL https://www.sciencedirect.com/science/article/pii/S1084804523000723.

Botyarov, M., Miller, E.E., 2022. Partitioning around medoids as a systematic approach to generative design solution space reduction. Res. Eng. 15, 100544. http://dx.doi.org/10.1016/j.rineng.2022.100544, URL https://www.sciencedirect.com/science/article/pii/S2590123022002146.

Bouke, M.A., Abdullah, A., ALshatebi, S.H., Abdullah, M.T., Atigh, H.E., 2023. An intelligent DDoS attack detection tree-based model using gini index feature selection method. Microprocess. Microsyst. 98, 104823. http://dx.doi.org/10.1016/j.micpro.2023.104823, URL https://www.sciencedirect.com/science/article/pii/S0141933123000698.

Bremer, L., den Nijs, S., de Groot, H.L., 2024. The energy efficiency gap and barriers to investments: Evidence from a firm survey in The Netherlands. Energy Econ. 133, 107498. http://dx.doi.org/10.1016/j.eneco.2024.107498, URL https://www.sciencedirect.com/science/article/pii/S0140988324002068.

Brusco, M.J., Steinley, D., 2007. A comparison of heuristic procedures for minimum within-cluster sums of squares partitioning. Psychometrika 72, 583–600.

Chen, M.-W., Chang, M.-S., Mao, Y., Hu, S., Kung, C.-C., 2023. Machine learning in the evaluation and prediction models of biochar application: A review. Sci. Prog. 106 (1), 00368504221148842.

Deloitte, 2016. Energy efficiency in Europe | Deloitte | Energy & Resources. https://www2.deloitte.com/global/en/pages/energy-and-resources/articles/energy-efficiency-in-europe.html, (Accessed 30 January 2022).

Doukas, H., 2018. On the appraisal of Triple-A energy efficiency investments. Energy Sour. B: Econ. Plan. Policy 13 (7), 320–327.

Doukas, H., Xidonas, P., Mastromichalakis, N., 2021. How successful are energy efficiency investments? A comparative analysis for classification & performance prediction. Comput. Econ. 1–20.

EEFIG, 2017. EEFIG UNDERWRITING TOOLKIT Value and risk appraisal for energy efficiency financing. https://eefig.ec.europa.eu/system/files/2020-11/EEFIG_Underwriting_Toolkit_June_2017.pdf, (Accessed 30 August 2023).

Geyer, P., Schlüter, A., Cisar, S., 2017. Application of clustering for the development of retrofit strategies for large building stocks. Adv. Eng. Inform. 31, 32–47. http://dx.doi.org/10.1016/j.aei.2016.02.001, URL https://www.sciencedirect.com/science/article/pii/S1474034616300167, Towards a new generation of the smart built environment.

Ghattas, B., Michel, P., Boyer, L., 2017. Clustering nominal data using unsupervised binary decision trees: Comparisons with the state of the art methods. Pattern Recognit. 67, 177–185. http://dx.doi.org/10.1016/j.patcog.2017.01.031, URL https://www.sciencedirect.com/science/article/pii/S0031320317300390.

Gillingham, K., Palmer, K., 2014. Bridging the energy efficiency gap: Policy insights from economic theory and empirical evidence. Rev. Environ. Econ. Policy.

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics 857–871.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. J. R. Stat. Soc. Ser. C (Appl. Stat.) 28 (1), 100–108.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Vol. 2, Springer.

Hill, D.R., 2019. Energy efficiency financing: A review of risks and uncertainties. In: Energy Challenges for the Next Decade, 16th IAEE European Conference. Vol. 29, pp. 249–262.

IEA, 2021. Global energy efficiency progress is recovering – but not quickly enough to meet international climate goals. https://www.iea.org/news/global-energy-efficiency-progress-is-recovering-but-not-quickly-enough-to-meet-international-climate-goals, (Accessed 28 February 2022).

Ienco, D., Pensa, R.G., Meo, R., 2012. From context to distance: Learning dissimilarity for categorical data clustering. ACM Trans. Knowl. Discov. Data (TKDD) 6 (1), 1–25.

Jebli, I., Belouadha, F.-Z., Kabbaj, M.I., Tilioua, A., 2021. Prediction of solar energy guided by pearson correlation using machine learning. Energy 224, 120109. http://dx.doi.org/10.1016/j.energy.2021.120109, URL https://www.sciencedirect.com/science/article/pii/S0360544221003583.

Jiang, M., Wang, J., Hu, L., He, Z., 2023. Random forest clustering for discrete sequences. Pattern Recognit. Lett. 174, 145–151. http://dx.doi.org/10.1016/j.patrec.2023.09.001, URL https://www.sciencedirect.com/science/article/pii/S0167865523002507.

Khan, W., Walker, S., Zeiler, W., 2023. A bottom-up framework for analysing city-scale energy data using high dimension reduction techniques. Sustainable Cities Soc. 89, 104323. http://dx.doi.org/10.1016/j.scs.2022.104323, URL https://www.sciencedirect.com/science/article/pii/S2210670722006278.

Kleanthis, N., Koutsandreas, D., Karakosta, C., Doukas, H., Flamos, A., 2022. Bridging the transparency gap in energy efficiency financing by co-designing an integrated assessment framework with involved actors. Energy Rep. 8, 9686–9699. http://dx.doi.org/10.1016/j.egyr.2022.07.066, URL https://www.sciencedirect.com/science/article/pii/S2352484722013312.

Koutsandreas, D., 2023. Does complexity compensate for accuracy in annual final energy demand forecasting? A multi-methods case study in G7 countries. In: 2023 19th International Conference on the European Energy Market. EEM, pp. 1–7. http://dx.doi.org/10.1109/EEM58374.2023.10161829.

Koutsandreas, D., Kleanthis, N., Flamos, A., Karakosta, C., Doukas, H., 2022. Risks and mitigation strategies in energy efficiency financing: A systematic literature review. Energy Rep. 8, 1789–1802. http://dx.doi.org/10.1016/j.egyr.2022.01.006, URL https://www.sciencedirect.com/science/article/pii/S2352484722000063.

Lahmiri, S., 2024. Fossil energy market price prediction by using machine learning with optimal hyper-parameters: A comparative study. Resour. Policy 92, 105008. http://dx.doi.org/10.1016/j.resourpol.2024.105008, URL https://www.sciencedirect.com/science/article/pii/S0301420724003751.

Li, S., Leng, Y., Abed, A.M., Dutta, A.K., Ganiyeva, O., Fouad, Y., 2024. Waste-to-energy poly-generation scheme for hydrogen/freshwater/power/oxygen/heating capacity production; optimized by regression machine learning algorithms. Process Saf. Environ. Prot. 187, 876–891. http://dx.doi.org/10.1016/j.psep.2024.04.118, URL https://www.sciencedirect.com/science/article/pii/S0957582024004749.

Li, P.-H., Pye, S., Keppo, I., 2020. Using clustering algorithms to characterise uncertain long-term decarbonisation pathways. Appl. Energy 268, 114947. http://dx.doi.org/10.1016/j.apenergy.2020.114947, URL https://www.sciencedirect.com/science/article/pii/S0306261920304591.

Liao, N., He, Y., 2018. Exploring the effects of influencing factors on energy efficiency in industrial sector using cluster analysis and panel regression model. Energy 158, 782–795. http://dx.doi.org/10.1016/j.energy.2018.06.049, URL https://www.sciencedirect.com/science/article/pii/S0360544218311149.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22, URL https://CRAN.R-project.org/doc/Rnews/.

Liu, G., Ji, F., Sun, W., Sun, L., 2023. Optimization design of short-circuit test platform for the distribution network of integrated power system based on improved K-means clustering. Energy Rep. 9, 716–726. http://dx.doi.org/10.1016/j.egyr.2023.04.319, URL https://www.sciencedirect.com/science/article/pii/S2352484723006868, 2022 The 3rd International Conference on Power Engineering.

Loureiro, T., Gil, M., Desmaris, R., Andaloro, A., Karakosta, C., Plesser, S., 2020. De-risking energy efficiency investments through innovation. Proceedings 65 (1), http://dx.doi.org/10.3390/proceedings2020065003, URL https://www.mdpi.com/2504-3900/65/1/3.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2021. cluster: Cluster Analysis Basics and Extensions. R package version 2.1. 2—For new features, see the'Changelog'file (in the package source). Vol. 1, R Package Version R Foundation for Statistical Computing, Vienna, Austria, p. 56.

Mehedi Hassan, M., Mollick, S., Yasmin, F., 2022. An unsupervised cluster-based feature grouping model for early diabetes detection. Healthc. Anal. 2, 100112. http://dx.doi.org/10.1016/j.health.2022.100112, URL https://www.sciencedirect.com/science/article/pii/S2772442522000521.

Metzig, C., Gould, M., Noronha, R., Abbey, R., Sandler, M., Colijn, C., 2020. Classification of origin with feature selection and network construction for folk tunes. Pattern Recognit. Lett. 133, 356–364. http://dx.doi.org/10.1016/j.patrec.2020.03.023, URL https://www.sciencedirect.com/science/article/pii/S016786552030101X.

Mexis, F.D., Papapostolou, A., Karakosta, C., Sarmas, E., Koutsandreas, D., Doukas, H., 2021. Leveraging energy efficiency investments: An innovative web-based benchmarking tool. Adv. Sci. Technol. Eng. Syst. J. 6 (5), 237–248.

Moksnes, N., Rozenberg, J., Broad, O., Taliotis, C., Howells, M., Rogner, H., 2019. Determinants of energy futures—a scenario discovery method applied to cost and carbon emission futures for South American electricity infrastructure. Environ. Res. Commun. 1 (2), 025001. http://dx.doi.org/10.1088/2515-7620/ab06de.

Papadopoulos, S., Bonczak, B., Kontokosta, C.E., 2018. Pattern recognition in building energy performance over time using energy benchmarking data. Appl. Energy 221, 576–586. http://dx.doi.org/10.1016/j.apenergy.2018.03.079, URL https://www.sciencedirect.com/science/article/pii/S0306261918304070.

Ping, Y., Li, H., Hao, B., Guo, C., Wang, B., 2024. Beyond k-Means++: Towards better cluster exploration with geometrical information. Pattern Recognit. 146, 110036. http://dx.doi.org/10.1016/j.patcog.2023.110036, URL https://www.sciencedirect.com/science/article/pii/S0031320323007331.

Pye, S., Li, P.-H., Keppo, I., O'Gallachoir, B., 2019. Technology interdependency in the United Kingdom's low carbon energy transition. Energy Strategy Rev. 24, 314–330. http://dx.doi.org/10.1016/j.esr.2019.04.002, URL https://www.sciencedirect.com/science/article/pii/S2211467X1930029X.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. Vienna, Austria, URL http://www.R-project.org/, ISBN 3-900051-07-0.

Rezessy, S., Bertoldi, P., 2010. Financing Energy Efficiency: Forging the Link Between Financing and Project Implementation. Vol. 152, Report prepared by the Joint Research Centre of the European Commission.

Rubino, A., 2017. Energy efficiency: Governance in the EU. Nat. Energy 2 (6), 17097.

Sarmas, E., Spiliotis, E., Marinakis, V., Koutselis, T., Doukas, H., 2022. A meta-learning classification model for supporting decisions on energy efficiency investments. Energy Build. 258, 111836. http://dx.doi.org/10.1016/j.enbuild.2022.111836, URL https://www.sciencedirect.com/science/article/pii/S037877882200007X.

Shivakumar, A., Alfstad, T., Niet, T., 2021. A clustering approach to improve spatial representation in water-energy-food models. Environ. Res. Lett. 16 (11), 114027. http://dx.doi.org/10.1088/1748-9326/ac2ce9.

Shvili, J., 2021. Regions Of Europe. https://www.worldatlas.com/articles/the-four-european-regions-as-defined-by-the-united-nations-geoscheme-for-europe.html, (Accessed 10 July 2023).

Sorrell, S., Gatersleben, B., Druckman, A., 2020. The limits of energy sufficiency: A review of the evidence for rebound effects and negative spillovers from behavioural change. Energy Res. Soc. Sci. 64, 101439. http://dx.doi.org/10.1016/j.erss.2020.101439, URL https://www.sciencedirect.com/science/article/pii/S2214629620300165.

Stevens, D., Adan, H., Brounen, D., de Coo, W., Fuerst, F., Kavarnou, D., Singh, R., 2019. Risks and uncertainties associated with residential energy efficiency investments. Real Estate Finance 35 (4), 249–262.

Uti, M.N., Md Din, A.H., Yusof, N., Yaakob, O., 2023. A spatial-temporal clustering for low ocean renewable energy resources using K-means clustering. Renew. Energy 219, 119549. http://dx.doi.org/10.1016/j.renene.2023.119549, URL https://www.sciencedirect.com/science/article/pii/S0960148123014647.

van der Maaten, L., 2014. Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. 15, 3221–3245.

van der Maaten, L., Hinton, G., 2008. Visualizing high-dimensional data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Wang, Q., Lee, B.D., Augenbroe, G., Paredis, C.J., 2017. An application of normative decision theory to the valuation of energy efficiency investments under uncertainty. Autom. Constr. 73, 78–87. http://dx.doi.org/10.1016/j.autcon.2016.09.005, URL https://www.sciencedirect.com/science/article/pii/S0926580516302242.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, http://dx.doi.org/10.1111/j.1467-985X.2010.00676_9.x, URL https://ggplot2.tidyverse.org.

Yaro, A.S., Maly, F., Prazak, P., 2023. Outlier detection in time-series receive signal strength observation using Z-score method with s n scale estimator for indoor localization. Appl. Sci. 13 (6), 3900.