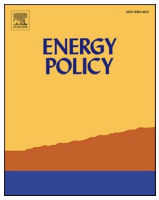
Machine Learning Framework for Evaluating Energy Performance Certificate (EPC) Effectiveness in Real Estate: The Case Study of Turin’s Private Residential Market

Energy Policy 198 (2025) 114407

Contents lists available at ScienceDirect

Energy Policy

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



Machine learning framework for evaluating energy performance certificate (EPC) effectiveness in real estate: A case study of Turin's private residential market

Federico Dell'Anna

Interuniversity Department of Regional and Urban Studies and Planning (DIST), Politecnico di Torino, Viale Mattioli 39, 10125, Turin, Italy

ARTICLE INFO

Keywords:
Deep learning (DL)
Hedonic pricing method (HPM)
SHapley additive exPlanations (SHAP)
Green premium
Real estate market segmentation
Mass appraisal

ABSTRACT

The Energy Performance Certificate (EPC) is a key tool for advancing building energy efficiency across Europe. By offering standardized information on a property's energy use, it shapes buyer and tenant preferences, influencing property values. This data-driven policy analysis assesses the EPC's effectiveness.

To assess the impact of EPC on property prices in Turin, Italy, a comprehensive machine learning (ML) framework is employed. This framework includes unsupervised hierarchical clustering and supervised algorithms including Artificial Neural Networks (ANN), k-Nearest Neighbors (k-NN), Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting Machine (GBM). These techniques facilitate an in-depth analysis of the complex relationships between EPC ratings and property prices.

Furthermore, the integration of eXplainable Artificial Intelligence (XAI) enhances the transparency of these models, providing clear insights into how EPC ratings affect prices across different property sub-markets. By demystifying the decision-making processes of complex algorithms, this approach makes the findings more accessible to stakeholders.

The flexibility of this framework suggests that it can be applied to other European contexts, offering a valuable tool for policymakers aiming to craft more effective energy efficiency strategies.

1. Introduction

Due to significant energy consumption in the building sector, the EU introduced the Energy Performance Certificate (EPC) to benchmark energy efficiency and influence regulations (European Commission, 2002). The revised Energy Performance of Buildings Directive (EPBD, EU/2024/1275) standardizes EPC across the EU by integrating indicators for both energy use and greenhouse gas emissions. This standardization benefits building owners and tenants with clearer energy data that can enhance property value and reduce costs. Financial institutions gain consistent metrics for assessing risks and opportunities in green financing (Brown et al., 2019; Raushan et al., 2024), while public authorities find it simplifies compliance monitoring and supports climate targets (European Commission, 2024).

The directive introduces a common A-G rating scale, with 'A' representing zero-emission buildings. Some Member States (MSs) may add an 'A+' rating for properties exceeding zero-emission standards. It mandates the inclusion of EPC in key property transactions, increasing their influence in the real estate market. Additionally, the EPBD requires national energy performance databases and building renovation passports to enhance transparency and accountability for property owners (European Commission, 2024).

Sustainability is increasingly shaping the field of property appraisal, with growing research on how energy efficiency influences property valuations and how EPC policies affect consumer preferences (Owen et al., 2023; Schuitema et al., 2020; D'Alpaos and Bragolusi, 2018).

The hedonic pricing method (HPM) is widely used to assess the impact of EPC and international sustainability certifications like LEED, BREEAM, and Green Mark on property values. Studies consistently demonstrate a positive relationship between higher certification levels and increased property values, highlighting the economic benefits of enhanced energy performance and sustainable building practices (Barreca et al., 2021; Bisello et al., 2020; Cajias et al., 2019; Chegut et al., 2011; Copiello and Donati, 2021; Costa et al., 2018; Dell'Anna and Bottero, 2021; Deng et al., 2012; Fuerst et al., 2016b; Taltavull de La Paz et al., 2019; Tsai, 2022).

However, HPM has limitations, including its assumption of linearity and susceptibility to omitted variable bias. In contrast, machine learning

E-mail address: federico.dellanna@polito.it.

https://doi.org/10.1016/j.enpol.2024.114407
Received 25 June 2024; Received in revised form 18 October 2024; Accepted 31 October 2024
Available online 10 December 2024

(ML) algorithms offer advanced solutions by handling large datasets and capturing complex, non-linear relationships between features and property values. This capability is particularly relevant for EPC ratings, which often have non-linear effects on market values (Cajias, 2021). Additionally, variations across different housing markets and geographic regions present challenges in evaluating EPC impacts, necessitating a more spatially aware approach (Dell'Anna, 2022; Marmolejo-Duarte and Chen, 2019; McCord et al., 2020).

To address these challenges, this study proposes a flexible framework suitable for diverse European contexts, employing a dual approach to capture spatial complexities and non-linearity. Hierarchical clustering segments urban areas based on real estate characteristics, enabling detailed analysis of EPC impacts within each segment. After segmentation, both the HPM and advanced ML techniques, including Artificial Neural Networks (ANN), Support Vector Regression (SVR), k-Nearest Neighbors (k-NN), Random Forest (RF), and Gradient Boosting Machine (GBM), are used to predict the effects of EPC ratings and other property attributes like location, size, age, and building type on property values. Addiotionally, eXplainable Artificial Intelligence (XAI) tools are employed to interpret the models predictions, providing insights into the influence of individual features on property value outcomes.

Applied to Turin, Italy, where the impact of EPC on property prices was previously analyzed using traditional methods, this framework incorporates advanced ML and deep learning (DL) techniques to provide a more comprehensive understanding of how EPC ratings influence real estate prices, validating and refining earlier findings. This application enhances the credibility of ML techniques and demonstrates the framework's adaptability to different European markets.

The paper is structured as follows: Section 2 reviews econometric and ML approaches used in real estate market analysis; Section 3 outlines the proposed methodological framework; Section 4 presents the case study and results; and Section 5 discusses the implications of the findings for data-driven energy policy analysis.

2. Modelling the housing market through mass appraisal

2.1. Introduction to machine learning in real estate

ML, a branch of artificial intelligence (AI), enables systems to learn from data, recognize patterns, and make decisions with minimal human intervention. In the real estate sector, ML and DL techniques offer powerful tools for analyzing complex datasets, enhancing decision-making in property valuation, market analysis, and investment strategies.

ML applications in real estate include supervised learning, classification, and clustering, each suited to specific tasks. Supervised learning models, particularly regression algorithms, are widely used in property valuation. They leverage historical data to uncover relationships between property characteristics and accurately predict prices (Ja'afar et al., 2021; Valier and Micelli, 2020).

Unsupervised learning techniques, such as k-means clustering, hierarchical clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), are essential for market segmentation. These methods identify homogeneous submarkets or clusters of properties with similar characteristics, improving the precision of valuation models by accounting for local market variations (Al-Qawasmi, 2022; Mete and Yomralioglu, 2023; Murtagh and Contreras, 2012; Skovajsa, 2023).

DL algorithms like ANN, Deep Neural Networks (DNN), and Convolutional Neural Networks (CNN) offer enhanced capabilities for processing complex, high-dimensional data. DL helps capture temporal trends in real estate prices and demand, enabling more precise market forecasting. For example, Recurrent Neural Networks (RNN) are particularly useful for understanding temporal sequences in market behavior (Al-Qawasmi, 2022).

Generative AI is also influencing real estate valuation by introducing novel approaches for model generation, simulation, and data augmentation. Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE) are gaining popularity for their ability to generate synthetic data, enhancing valuation models by simulating diverse market conditions. GAN create realistic datasets that capture the complexities of shifting market dynamics, contributing to more accurate property valuations (Lee, 2021; Zhao et al., 2023).

The following sections will explore both unsupervised and supervised machine learning, as well as deep learning techniques, to determine which methods are best suited to address key challenges in market segmentation and price prediction within the context of EPC analysis.

2.2. Spatial variability on property values and market segmentation

Traditional urban economic models focus on location as the main determinant of property values (Alonso, 1964). However, market segmentation considers a broader set of variables influencing housing markets, capturing the unique behaviors of different submarkets. While location remains critical, variations in neighborhood amenities, property characteristics like size and age, and socioeconomic factors lead to significant price differences within the same city (Can, 1992; Galster, 1996; Goodman and Thibodeau, 1998; Maclennan and Tu, 1996; Bottero et al., 2022; Thackway et al., 2022).

Real estate segmentation frequently uses administrative boundaries, ZIP codes, or census tracts, to define submarkets that help control for location-specific effects in HPM. However, these boundaries often fail to capture the complexity of local market dynamics (Bourassa et al., 1999; Palm, 1978; Schnare and Struyk, 1976). Goodman et al. (2003) advocated for a more flexible, data-driven approach to better reflect actual market variability.

The need for more refined segmentation is also evident in previous studies on EPC impacts. Marmolejo-Duarte and Chen (2019) found that in Barcelona, the influence of EPC ratings on property values varies by property type and socioeconomic conditions, with low EPC ratings decreasing values in lower-income areas, underscoring the necessity for targeted retrofitting policies. Marmolejo-Duarte et al. (2020) further showed that the negative impact of low EPC ratings is more pronounced in lower-quality, peripheral neighborhoods, acting as a quality differentiator in markets lacking basic amenities. Similarly, McCord et al. (2020) demonstrated that the effect of EPC ratings in Belfast is spatially heterogeneous, with some areas showing a premium for energy-efficient homes while others exhibited no significant difference, reflecting regional variations in consumer awareness. Dell'Anna, 2022 also noted variability in EPC impact in Turin, highlighting the need for localized energy policies. These findings underscore the importance of market segmentation for effectively analyzing EPC impacts and designing targeted, equitable energy efficiency policies.

Advances in spatial econometrics and ML have revolutionized market segmentation in real estate. Manganelli et al. (2014) and McCluskey and Borst (2011) demonstrated the effectiveness of Geographically Weighted Regression (GWR) in identifying submarkets by capturing spatial variations in property values, resulting in better predictive accuracy compared to global models. However, GWR faces challenges with large datasets, multicollinearity, and linear assumptions, which can oversimplify the complex dynamics of real estate markets.

In contrast, unsupervised machine learning techniques, such as k-means clustering, DBSCAN, Gaussian Mixture Models (GMM), Agglomerative Hierarchical Clustering, and Self-Organizing Maps (SOM), offer flexible approaches to market segmentation by processing multiple features and capturing both spatial and non-spatial characteristics (Bhagat et al., 2020; Gružauskas et al., 2021; Heidari et al., 2021; Napoli et al., 2017; Skovajsa, 2023).

K-means clustering is widely used for its efficiency and simplicity (Del Giudice et al., 2024; Bourassa et al., 1999; Gabrielli et al., 2019). Wiersma et al. (2022) combined k-means with Ward's method to classify German housing markets, while Kim and Irakoze (2022) used k-means to

cluster green-certified apartment submarkets before applying HPM for price predictions, demonstrating enhanced accuracy in predicting the green premium. However, k-means requires a predefined number of clusters, limiting its flexibility for diverse datasets.

DBSCAN excels at detecting clusters with arbitrary shapes and effectively handles noise in complex datasets (Birant and Kut, 2007; Ester et al., 1996). Liu et al. (2012) integrated spatial proximity and attribute similarity into DBSCAN using Delaunay triangulation, identifying clusters of varying shapes and densities. Unlike k-means, DBSCAN does not require a predefined number of clusters but is sensitive to parameter settings, which may limit its generalizability in highly heterogeneous markets.

Hierarchical clustering is well-suited for real estate segmentation as it captures the hierarchical structure of submarkets without requiring predefined boundaries (Skovajsa, 2023). For instance, Ara Aksoy and Irwin (2021) demonstrated its flexibility in handling mixed-mode data, integrating diverse property characteristics. It is more interpretable than k-means and can reveal deeper relationships. In a study on the UK residential market using hierarchical clustering, Konhäuser and Werner (2024) further enhance interpretability by incorporating eXplainable Artificial Intelligence (XAI) tools, such as SHapley Additive exPlanations (SHAP), Permutation Feature Importance (PFI), and Partial Dependence Plots (PDPs). The authors highlighted the value of XAI in helping policymakers better understand how EPC ratings influence property values, enabling more informed decisions and effective energy efficiency policies.

2.3. Limitations and advantages of price prediction methods

Big data has revolutionized our understanding of market dynamics by providing vast amounts of information, and econometric analyses offer a structured way to utilize this data (Rosen, 1974). The HPM is a widely used tool in real estate economics, estimating the implicit prices of property attributes based on observed market prices. While HPM is valuable for isolating the impact of factors like location, size, and energy efficiency, it has inherent limitations.

One major challenge is handling uncaptured variability. HPM assumes linear relationships, but real-world data often exhibit non-linear patterns (Cropper et al., 1988). In the context of EPC ratings, non-linearity has been addressed by creating dummy variables for each class or grouping them into categories (low, medium, high). Studies across European markets reveal clear non-linear relationships (Bisello et al., 2020; Micelli et al., 2023). In England and Wales, properties with poor energy efficiency experience substantial value increases when upgraded, but this impact lessens as homes reach higher EPC ratings (Fuerst et al., 2015; Jensen et al., 2016). Similarly, findings from Barcelona and Bucharest highlight that the green premium is most pronounced when moving properties from lower to average efficiency, with diminishing returns as energy performance improves further (Marmolejo Duarte, 2016; Taltavull et al., 2017). Non-linearities necessitate ML models to better capture complex dynamics. Treating EPC ratings as ordinal variables enables finer analysis of how changes in energy performance impact property values.

Another limitation of traditional econometric models is their vulnerability to omitted variable bias. Excluding relevant attributes like renovation quality, building condition, and architectural features can distort the estimated effects of energy efficiency on property prices (Fuerst et al., 2016a; Marmolejo-Duarte and Chen, 2022a, 2022b). In contrast, ML models like RF and GBM can automatically identify the importance of a wide range of variables, including those that may seem insignificant individually but interact to affect property values (Gao et al., 2022; Levantesi and Piscopo, 2020).

2.4. Willingness to pay for energy-efficient investigation

Factors influencing willingness to pay (WTP) for energy-efficient homes include economic, social, and environmental considerations. A primary driver is the perceived economic benefit from reduced energy bills; homeowners are willing to pay a premium for long-term savings when these are clearly communicated (Encinas et al., 2018). Comfort and co-benefits significantly shape WTP (Buso et al., 2017; Crespo Sánchez et al., 2021; Becchio et al., 2018). Studies show that comfort benefits, like improved thermal insulation and air quality, are valued comparably to direct energy savings (Banfi et al., 2008; Ferreira and Almeida, 2015; Ürge-Vorsatz et al., 2014). Post-pandemic, the importance of indoor comfort has grown, with households willing to invest in thermal and acoustic enhancements despite higher costs (Berto et al., 2023). Environmental awareness and social responsibility also drive WTP; individuals motivated by these factors are more likely to invest in green real estate projects, especially in areas with air quality concerns (Wang et al., 2015). Personalization and aesthetic enhancements tied to energy upgrades are additional drivers (Bottero et al., 2019). Aesthetic features significantly boost the marketability of energy-efficient buildings, increasing acceptance and demand (Aydin et al., 2019). Research into WTP across EPC ratings provides crucial insights into consumer behavior and market dynamics, aiding policymakers and developers in promoting energy-efficient homes that are sustainable and economically viable for diverse buyers.

While the HPM is widely used to analyze EPC impact on prices, the application of ML algorithms remains limited. Studies utilizing techniques like linear regression (LR), Decision Trees, RF, Ridge, and Lasso have shown that Decision Tree and RF algorithms perform better in predicting green building prices, highlighting ML growing role in real estate sustainability analysis. Jamil et al. (2020), Masrom et al. (2022), and Mohd et al. (2022) used ML algorithms to predict green building prices in Kuala Lumpur District (Malaysia). These studies highlighted the superior performance of Decision Tree and RF algorithms, demonstrating the growing role of ML in real estate sustainability analysis. In commercial real estate, EPC has been examined by Akhtyrska and Fuerst (2024) using ML techniques combined with Difference-in-Differences (DID) and panel data fixed effects. They showed that introducing Minimum Energy Efficiency Standards (MEES) in England and Wales reduced rents by 6–8% in affected office buildings.

Despite these advancements, significant gaps remain in using ML for analyzing EPC in the residential sector. Employing advanced ML models like GBM, SVM, and DL could improve EPC rating predictions and uncover complex data patterns overlooked by traditional methods. By integrating socio-economic factors, building characteristics, and geographic information, ML can provide a comprehensive view of the green premium in the real estate market.

A structured approach combining market segmentation, ML-supported price prediction, and eXplainable AI (XAI) can help MSs standardize EPC policy effectiveness assessments.

3. Methodology

3.1. Workflow overview

The methodological framework of this research is structured into four macro-phases to systematically analyze the real estate market and EPC ratings, as illustrated in Fig. 1. These macro-phases include: goal definition, data preparation, modeling, and evaluation and interpretation. Each phase involves multiple steps that ensure a comprehensive and systematic approach to analyzing the real estate market.

The first phase, goal definition, sets the research objectives that guide the entire process. In the second phase, data preparation, feature definition is conducted (Step 1) to identify variables that influence the market value of residential properties. In Step 2, data is collected from multiple sources, including real estate advertisements for property details, data for accessibility features, and socio-economic variables. Step 3 focuses on data cleaning and standardization, addressing missing values, detecting outliers, and ensuring data quality.

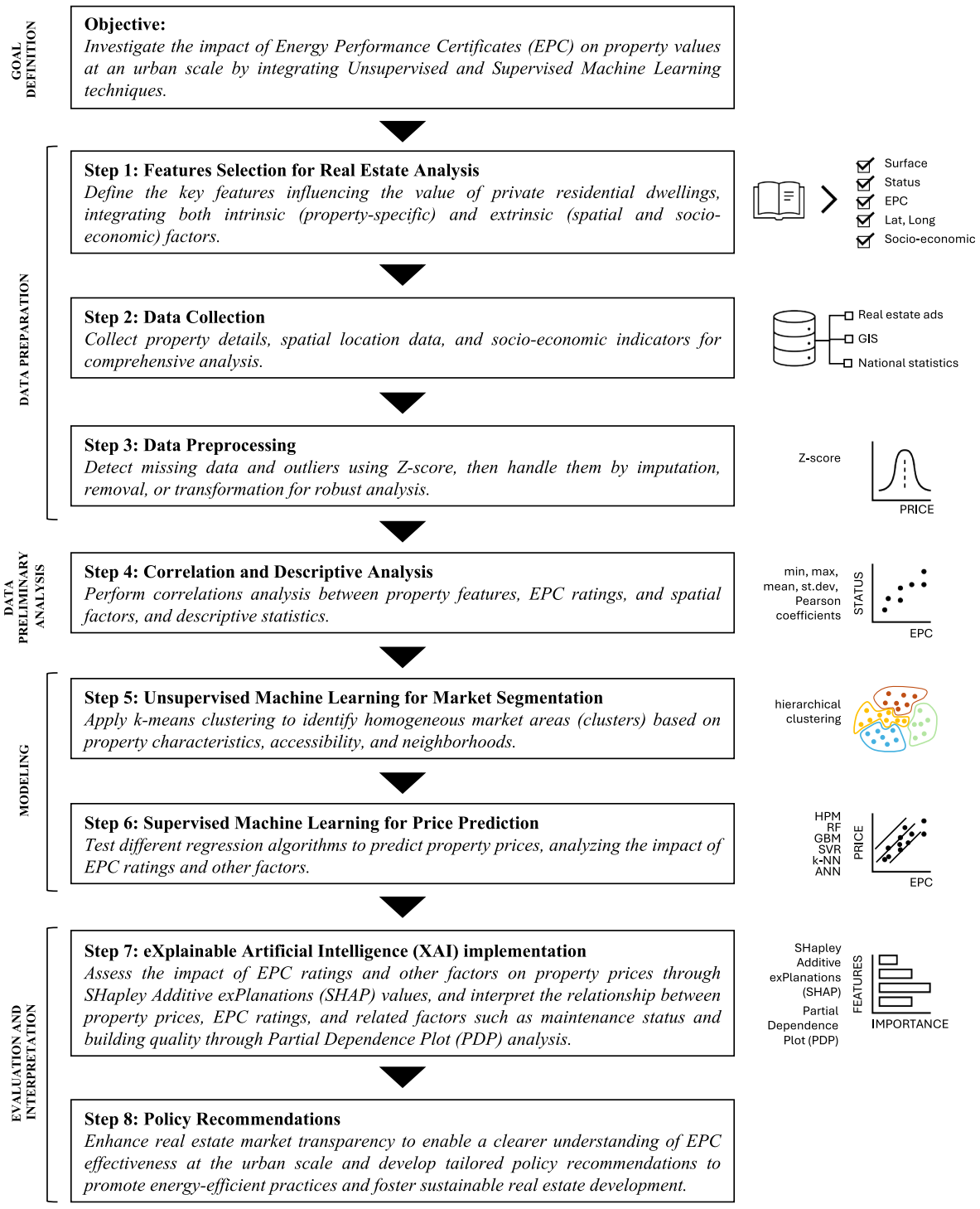


Fig. 1. Methodology workflow.

In the third phase, exploratory data analysis (Step 4) examines relationships within the dataset. Pearson correlation is used to explore interdependencies, and descriptive statistics provide an overview of data distribution.

The fourth phase, modeling, includes Steps 5 and 6. In Step 5, unsupervised ML, specifically hierarchical clustering, segments the real estate market into homogeneous zones based on intrinsic property characteristics and extrinsic factors to identify similar market behaviors. In Step 6, the HPM and advanced ML algorithms are used to predict property prices and estimate the impact of EPC ratings.

The final phase involves evaluation and interpretation. eXplainable Artificial Intelligence (XAI) techniques are applied in Step 7 to enhance interpretability. In Step 8, insights from the models are translated into policy recommendations.

3.2. Defining features influencing real estate value

The primary goal of Step 1 is to ensure that the selected variables are both relevant and representative of the complex factors influencing real estate dynamics. To enhance the robustness of the subsequent analyses, a literature-driven approach is recommended for variable selection, aligning the study with established theories and empirical evidence in

the field. This approach ensures that the chosen variables reflect the multifaceted nature of real estate markets, facilitating a comprehensive and accurate analysis. Generally, these variables include intrinsic factors such as price, size, age, technological features, and sustainability qualities, as well as extrinsic characteristics like location and socio-economic variables.

3.3. Data collection

The initial phase of constructing a detailed dataset for the study area involves gathering real estate data (Step 2). One key source is real estate advertisements, which provide extensive information on intrinsic property characteristics such as size, construction year, type, condition, and features. To further enrich the dataset, Geographic Information System (GIS) tools are used to integrate spatial factors, including distances to transportation networks and proximity to essential urban amenities like schools, green spaces, and public transport. This integration provides a more comprehensive understanding of the properties by considering their accessibility and proximity to important services. Additionally, the dataset is supplemented with demographic and socio-economic data from national statistical institutions.

3.4. Data preprocessing

Data preprocessing is a foundational stage in the ML workflow, significantly impacting the subsequent modeling process (Step 3). This phase involves data manipulation and cleaning procedures to improve data quality and suitability for modeling, ultimately enhancing the model's learning capacity and accuracy. Key preprocessing steps include:

- Outlier Detection and Handling: Identifying and addressing outliers is crucial because extreme values can distort analytical results and model performance. Detecting price outliers ensures that these anomalies do not disproportionately influence predictions. A common technique is using Z-scores to detect outliers, especially when data is approximately normally distributed. Z-scores measure how many standard deviations a data point is from the mean, with values greater than 3 or less than −3 typically considered outliers (Hodge and Austin, 2004).
- Categorical Variables: Converting categorical data into a numerical format suitable for ML algorithms, either as ordinal variables or through techniques like one-hot encoding.
- Target Variable Transformation: When the target variable (e.g., property prices) does not meet linear regression assumptions, applying transformations like the natural logarithm (ln) can help approximate a normal distribution, improving regression model performance.
- Standardization: Transforming continuous variables to have a mean of zero and a standard deviation of one (Z-score standardization) ensures uniformity among numerical features.

This rigorous data preprocessing ensures that the dataset is well-prepared for robust and accurate modeling in the subsequent analysis phases.

3.5. Correlation and descriptive analysis

Correlation analysis (Step 4) is used to elucidate relationships between variables, quantifying the extent of association between pairs and helping to identify potential predictors for the target variable. Pearson's correlation coefficients are applied to measure the strength and direction of linear relationships (Pearson, 1896). This analysis is crucial for guiding feature selection and model construction, ensuring that the most relevant variables are considered in subsequent modeling stages.

Descriptive analysis is another key step in understanding the dataset. It involves generating summary statistics, visualizations, and examining variable distributions to provide insights into central tendencies, variability, and the overall shape of the data.

3.6. Unsupervised machine learning algorithm for clustering

Hierarchical clustering, developed and popularized in the 1960s and 1970s, is a powerful unsupervised method for understanding relationships between groups of data. Although it lacks a single inventor, foundational concepts can be traced to Lance and Williams (1967), who contributed significantly by developing formulas to calculate distances between clusters. Hierarchical clustering begins by calculating distances between every pair of data points, using metrics like Euclidean or Manhattan distance. These distances are then used to construct a dendrogram, a tree-like structure that visually represents relationships among all data points, illustrating how clusters are progressively merged or divided (Murtagh and Contreras, 2012).

To form clusters, hierarchical clustering can use different linkage methods, which determine how distances between groups of data points are calculated during the clustering process. In this study, Ward's Method is selected. Ward's Method merges clusters by minimizing the increase in the within-cluster sum of squares, effectively reducing variance within clusters after each merge (Ward, 1963). This approach results in compact and homogeneous clusters, making it well-suited for achieving well-separated groupings and ensuring that the clusters are meaningful and interpretable.

A critical step in this process is determining the optimal number of clusters, often done using the elbow method (Shi et al., 2021). This method involves plotting within-cluster variance against the number of clusters and identifying the point where further reduction in variance starts to stabilize, forming an 'elbow' shape that indicates diminishing returns from adding more clusters.

Once the optimal number of clusters is determined, the dendrogram is cut at the appropriate level, providing segmentation of the dataset. The resulting clusters represent groups of data points with similar characteristics, offering structured insights into the inherent groupings within the data.

3.7. Regression algorithms for price prediction

Supervised machine learning algorithms play a critical role in predicting real estate property prices. An initial step involves creating training, validation, and testing sets as part of a systematic approach. Subsequent steps are specific to each algorithm, depending on the unique characteristics and requirements of the model (Step 6).

3.7.1. Train, validation and test datasets

Data is typically divided into three sets: a training set to fit the model, a validation set for hyperparameter tuning, and a test set to evaluate model performance. However, in many cases, this can be simplified to just training and test sets, particularly when the dataset is small or computational resources are limited (Vabalas et al., 2019).

To further enhance model evaluation, cross-validation is often employed to mitigate overfitting and provide a comprehensive assessment of model performance (Cawley and Talbot, 2010). In k-fold cross-validation, the dataset is split into 'k' equal parts, with each part used as validation data once while the others serve as training data. Averaging results across all 'k' iterations provides an overall performance estimate. Typically, 'k' ranges from 5 to 10 for larger datasets, ensuring that every segment of data is used for validation and resulting in a more reliable assessment of the model.

Hyperparameter tuning is an essential part of model development, such as determining the number of trees in a RF or the learning rate in a GBM. Random search is often used for hyperparameter optimization, testing different parameter combinations to find the configuration that minimizes validation error (Bergstra et al., 2012).

Proper selection of hyperparameters is crucial for model accuracy, as improper tuning can lead to overfitting or underfitting. The final hyperparameters are chosen based on cross-validation results to balance accuracy with generalizability.

To further refine the model and reduce overfitting, backward feature selection is applied, which is particularly effective for models like k-NN and ANN. By reducing the feature space, irrelevant or noisy variables are eliminated, which enhances the generalization of k-NN's distance-based predictions and simplifies ANN architectures. Combined with careful hyperparameter tuning, this process ensures optimal model performance while maintaining robustness against overfitting.

3.7.2. Linear regression model

Linear regression (LR) is a foundational and widely used algorithm within the context of HPM. It offers a straightforward and interpretable approach to modeling linear relationships between input features and a target variable, such as house prices (Lancaster, 1966; Rosen, 1974).

Linear regression performs well when the relationship between input features and house prices approximates linearity. This relationship is expressed through a linear equation (Eq. (1)):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (1)$$

where y represents the predicted house prices, β_0 is the intercept term, $\beta_1, \ldots, \beta_n$ are the coefficients associated with each input feature ($x_1, x_2, \ldots, x_n$), indicating the strength and direction of their influence on house prices[1,2]. This simple yet powerful method allows for the easy interpretation of how each feature contributes to the overall prediction, making it a valuable tool in real estate market analysis.

3.7.3. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN), inspired by biological neural networks, provide a complex, multi-layered modeling approach (McCulloch and Pitts, 1943). They are effective at processing large datasets, making them suitable for the intricate dynamics of real estate markets, where they can detect nuanced patterns beyond the scope of traditional methods. Effective use of ANN requires careful hyperparameter tuning, including choosing the number of layers, neurons, learning rate, and regularization techniques to prevent overfitting while capturing complex data relationships.[3] This involves adjusting hidden layers and decay rates to strike a balance between model complexity and generalizability.

3.7.4. K-Nearest Neighbors (k-NN)

The k-Nearest Neighbors (k-NN) algorithm estimates property values by averaging the prices of the 'k' most similar properties based on feature similarity (Altman, 1992; Cover and Hart, 1967). This method is particularly effective in uniform, densely populated areas and assumes that properties with similar features have comparable values (Yağmur et al., 2023). Choosing the optimal 'k' is crucial: a small 'k' may make the model overly sensitive to noise, while a large 'k' can smooth the results excessively. Similarity is measured using metrics such as Euclidean, Manhattan, or Minkowski distances, with Euclidean being the default. Feature weighting can vary from uniform to distance-based, where closer neighbors have more influence, affecting the model's accuracy.[4]

3.7.5. Support Vector Regression (SVR)

Support Vector Regression (SVR) adapts Support Vector Machine principles for regression tasks, offering a precise approach by fitting data within an epsilon margin to handle variability (Cortes and Vapnik, 1995). SVR is particularly effective for predicting property values, as it maintains prediction tolerance despite market volatility. Parameter selection, such as setting the epsilon value (ε), is crucial for controlling sensitivity; smaller ε values increase the model's responsiveness to price changes. SVR's versatility lies in its use of different kernel functions (linear, polynomial, radial basis function - RBF, sigmoid), with the RBF kernel often preferred for capturing non-linear relationships in property features. Adjusting the gamma parameter (γ) helps fine-tune the model to achieve an optimal balance between underfitting and overfitting.[5]

3.7.6. Random Forest (RF)

Random Forest (RF) consists of multiple decision trees, each trained on random subsets of features, which promotes diverse predictions and reduces the risk of overfitting. The final output is determined by aggregating the predictions from all the trees, typically through averaging (for regression) or voting (for classification) (Breiman, 2001).

RF's feature importance metric is particularly valuable, as it highlights the attributes that have the most significant influence on price variations. A key hyperparameter to tune in RF is the number of features randomly sampled as candidates at each split, which can be optimized through a random search within a specified range to improve model performance.

By combining the outputs of numerous decision trees, RF provides robust predictions and insights, making it a powerful tool for real estate market analysis. Its ability to handle large datasets with numerous features and its resistance to overfitting make it highly effective for modeling complex relationships in property values.[6]

3.7.7. Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) is a powerful, decision tree-based algorithm widely used for predictive tasks across various fields (Ding et al., 2022). It improves prediction accuracy by sequentially correcting errors from previous models, making it effective for a wide range of data types and distributions. GBM's ability to adapt to different loss functions allows it to address both classification and regression challenges, making it particularly useful for predicting outcomes such as disease classification or hospital stay durations in healthcare.

GBM incorporates several features to prevent overfitting, including data subsampling and limiting decision tree complexity, which help maintain model reliability. One of the significant advantages of GBM is its ability to provide insights into the factors most influencing outcomes, offering valuable guidance for decision-making, such as in real estate market analysis. Tuning GBM involves adjusting key hyperparameters, such as the depth of trees, number of trees, learning rate (shrinkage), and minimum observations in nodes, to optimize model performance.[7]

3.7.8. Validation approach

The validation approach for regression analysis and machine learning models involves a series of essential steps to assess their

[1] In this study, a semi-logarithmic linear regression model is employed where the dependent variable (price) is expressed in natural logarithms, allowing coefficients to represent the percentage change in price for a one-unit change in each independent variable, which aids in interpreting effects in percentage terms and is particularly useful for data exhibiting exponential growth patterns or heteroscedasticity.

[2] The 'lm' function, part of RStudio's base package, is used for performing linear regression (LR) analyses.

[3] The 'nnet' package provides functions for modeling Artificial Neural Networks (ANN).

[4] The 'clas' package offers the 'knn' function for implementing the k-Nearest Neighbors (k-NN) algorithm.

[5] The 'e1071' library provides functions for Support Vector Regression (SVR).

[6] The 'randomForest' package is used to implement the Random Forest (RF) model in Rstudio.

[7] The 'gbm' package is used to implement the Gradient Boosting Model (GBM) in RStudio.

performance and generalization capabilities. Three commonly used metrics for this purpose are RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and R-squared (R^2).

RMSE calculates the square root of the average of the squared differences between the predicted ($\hat{y}_i$) and actual (y_i) values, as following equation (Eq. (2)):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

It provides a measure of the average magnitude of prediction errors. RMSE with lower values indicates better model fit.

MAE computes the average of the absolute differences between predicted ($\hat{y}_i$) and actual (y_i) values, as following equation (Eq. (3)):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

It is robust to outliers since it doesn't square the errors. MAE with lower values indicates better model fit.

R-squared measures the proportion of the variance in the dependent variable (y) that is explained by the model. $\hat{y}_i$ represents predicted values, $\bar{y}$ is the mean of the observed values, and y_i represents actual values (Eq. (4)).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

R^2 ranges from 0 to 1, with higher values indicating a better fit of the model to the data.[8]

3.8. Results interpretation with eXplainable artificial intelligence (XAI)

In the European regulatory context, Trustworthy AI is formalized in the European Union's guidelines for reliable artificial intelligence, emphasizing transparency, accountability, and safety (European Commission, 2019). These principles are particularly critical in ML and DL, where model complexity can often be overwhelming for non-technical stakeholders, limiting their use in decision-making due to a lack of transparency. To address this challenge, eXplainable Artificial Intelligence (XAI) techniques are essential for interpreting and visualizing model outputs. Key XAI methods include ranking variables by importance using Shapley Additive Explanations (SHAP) values and utilizing Partial Dependence Plots (PDPs) to visualize the effects of variables on model outcomes (Step 7).

SHAP values are used to interpret model predictions by assigning each feature an importance value based on its contribution to the prediction (Fryer et al., 2020; Gao et al., 2022; Lundberg and Lee, 2017). The SHAP beeswarm plot visually summarizes the importance and impact of features, with each point representing a feature's contribution to an individual prediction. The x-axis ranks features from least to most important, while the y-axis displays SHAP values, indicating the magnitude and direction of their impact. Feature values are color-coded to indicate the relationship between the variable value and the magnitude of its impact on the prediction.

PDPs provide additional insights by illustrating how specific features influence predictions, capturing non-linear relationships and interactions. PDPs are calculated by averaging model predictions across different fixed values of a feature, isolating its effect while accounting for the influence of other features.[9] Together, SHAP and PDPs simplify complex model behavior into understandable visuals, enhancing trust and accessibility for stakeholders by translating intricate analyses into clear and actionable insights.

4. Case study: the city of Turin

4.1. Turin's architectural and energy legacy

Turin, in Northern Italy, is an ideal focus for studying energy performance in urban environments due to its diverse and historically rich building stock. The city's architecture, shaped over different periods of development, presents a mix of construction styles and standards, offering insights into how building characteristics impact energy efficiency and property value.

Around 60% of Turin's buildings were constructed before modern energy efficiency regulations, particularly from the pre-1945 period through the late 20th century. Many of these older structures lack proper thermal insulation and efficient heating, resulting in significant variability in energy performance. This makes Turin a valuable case study for exploring the relationship between building age, design, and energy efficiency, as well as for exploring localized energy policy strategies.

Turin's neighborhoods also reflect distinct socio-economic contexts, with varying building characteristics influencing both energy performance and market value.

Previous studies often used linear models to evaluate the effectiveness of EPC in Turin, which provided valuable but limited insights due to the assumption of linear relationships between variables (Bottero et al., 2018; Fregonara et al., 2014). The heterogeneous architectural landscape of Turin underscores the necessity for advanced methodologies. More advanced methods, such as ML, are better suited to address the non-linear relationships between energy performance, market dynamics, and neighborhood-specific factors, providing deeper insights into the effects of energy efficiency improvements.

4.2. Data selection and source

This study focuses on private residential apartment listings in Turin, representing the city's predominant residential architecture to evaluate the impact of EPC on the real estate market effectively, while also providing spatial insights across the municipal area. The year 2021 was selected due to data availability constraints and the unique characteristics of real estate listings from that period, allowing for a focused analysis of the market's post-COVID-19 evolution, including shifts in buyer behavior and housing preferences.

Most empirical studies on housing markets use sale prices as the dependent variable, as they are generally more accurate than valuations by owners or real estate agents, thereby reducing potential bias. However, Michelangeli (2008) noted that accessing final transaction data and detailed EPC information in Italy is challenging due to privacy and access restrictions. Therefore, data for this study was sourced from Immobiliare.it, Italy's leading real estate listings platform, which provides comprehensive data from both agencies and private sellers. Previous research, such as Curto et al. (2012), highlighted transparency issues in the Italian market and concluded that asking prices serve as reasonable estimators of market trends. While transaction prices would be ideal, asking prices still offer a meaningful understanding of market

[8] The 'Metrics' package in RStudio is used to compute various performance metrics, which are crucial for evaluating the effectiveness of machine learning models.

[9] The 'iml' package in RStudio is utilized for SHAP value calculations. PDPs are calculated using the 'pdp' package, and 'ggplot2' for data visualization. Calculating SHAP values and PDPs on a hold-out set has minimal impact, as model behavior is consistent across training and test sets. Using the training set for SHAP enhances representation fidelity by leveraging a "larger set" (Lundberg, 2018).

conditions.

4.3. Descriptive analysis

After pre-processing the data and handling outliers, the initial sample of 3314 observations was reduced to 2873 (Fig. 2). This reduction was due to the removal of 366 records with missing data and 16 outliers identified using Z-scores. 59 apartments located in the Turin Hills area were removed due to their distinct market characteristics. This rigorous data refinement ensures the integrity of the analysis by mitigating the influence of anomalies.

The selected variables, along with relevant summary statistics (Table A1), are briefly explained to enhance understanding. The variable 'ln_price' represents the logarithmically transformed values of real estate listings. The average asking price is approximately €1886/m^2, while the average property value based on 2021 data from the Agenzia delle Entrate is €1765/m^2 (Mazzitelli and Moine, 2022). Asking prices tend to be higher than final sale prices, as they do not account for negotiations that often lead to lower transaction values. This difference, approximately 6.85%, reflects the common practice where asking prices serve as an initial offer, typically reduced through bargaining before the sale is finalized.

The dataset's surface areas range from 28 to 800 m^2, with an average of 87.23 m^2, reflecting the compact nature of urban spaces in Italian cities. Approximately 75% of the properties are equipped with elevators, a common feature in urban settings. Additional variables include floor level, number of bathrooms, property category (ranging from economical to luxury), maintenance condition, and garage availability.

The average construction year of 1950 indicates that many properties were built during the mid-20th century, a period of significant growth for Turin. Distance variables are also included, measuring Euclidean distances to key locations such as the city center, universities, parks, subway stations, and bike paths.

A key focus of this study is the EPC variable, which rates properties on a scale from G (1 - least efficient) to A (7 - most efficient). The average EPC rating is 2.86, highlighting the prevalence of suboptimal energy performance in Turin's housing stock and emphasizing the need to examine how EPC ratings affect property values to inform and promote effective energy efficiency policies.

The dataset also includes socio-economic variables from the 2021 ISTAT census, such as population density, percentage of foreign residents, age structure, employment rate, education level, and income inequality (Gini Index), providing a comprehensive view of Turin's demographic and economic landscape. These factors are essential for understanding the broader socio-economic context influencing property pricing and market dynamics.

Additionally, the inclusion of latitude and longitude facilitates advanced spatial analysis, enabling the capture of complex geographic relationships. These variables enhance the ability of machine learning models, such as k-means clustering, to learn spatial patterns more

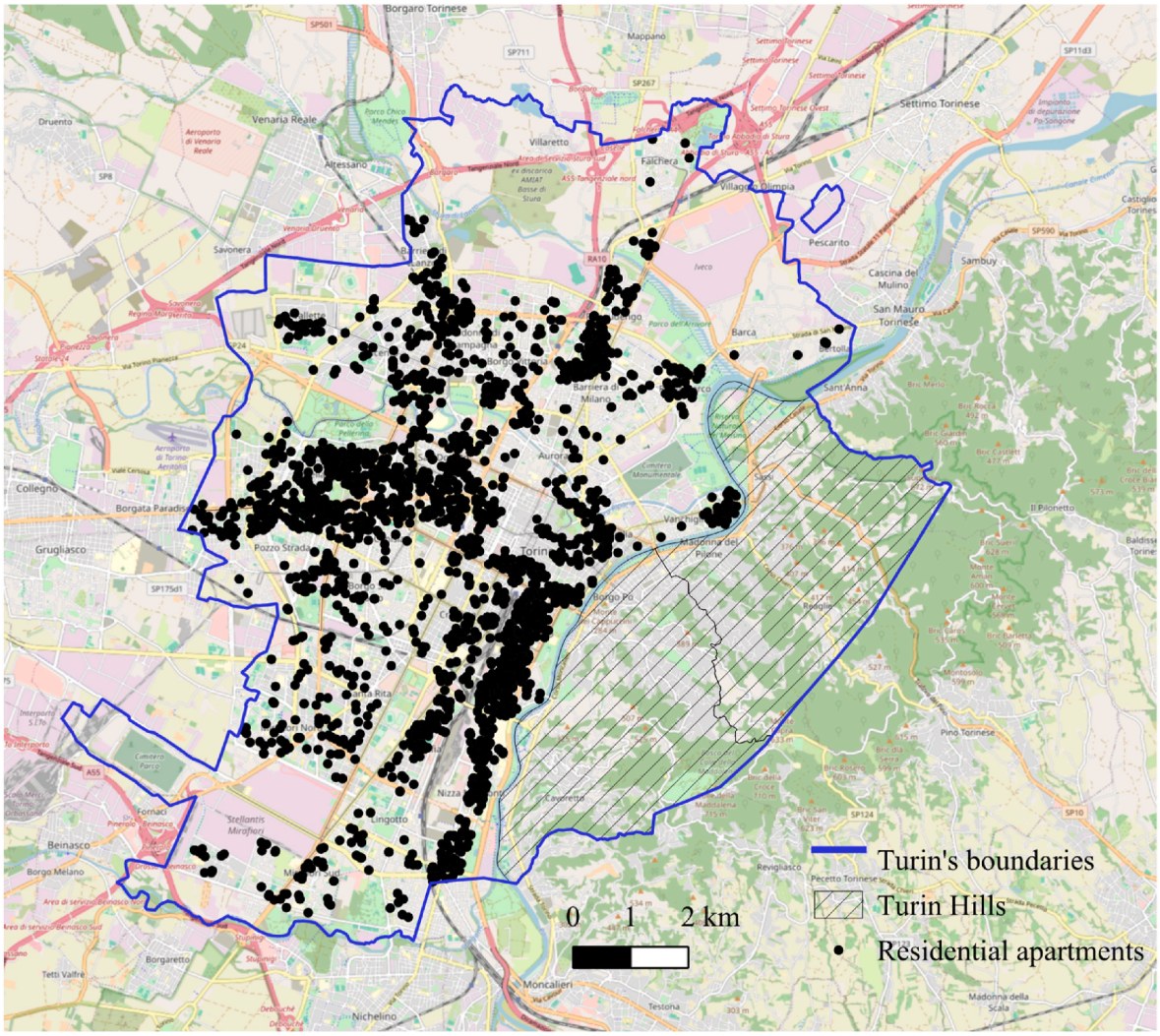


Fig. 2. Observations distribution in the Turin's municipal context.

effectively.

The dataset, enriched with detailed property information, socio-economic, and geographic variables, provides a robust foundation for analyzing Turin's residential real estate market.

4.4. Correlation analysis

The Pearson correlation analysis reveals key relationships between real estate attributes (Fig. 3). Surface area is positively correlated with the number of bathrooms (0.691) and, to a lesser extent, with property type (0.325), suggesting that larger properties often have more bathrooms and are categorized as higher-quality or luxury.

The EPC variable shows a moderate correlation with property condition (0.465), indicating that newer or renovated properties tend to have better energy ratings. EPC also correlates positively with property type, suggesting that energy-efficient buildings are more likely to belong to higher-quality categories. Additionally, property condition has a slight positive correlation with the year of construction (0.146), implying that newer properties tend to be better maintained, supporting the internal consistency of the dataset.

Geographically, older properties are generally closer to the city center and universities, as indicated by the positive correlations between year of construction and distance to the city center (0.427) and university campuses (0.315).

Socio-economic variables, such as population density, percentage of foreign residents, old age index, employment rate, education level, and Gini index, reveal complex interrelationships that reflect the city's socio-economic dynamics. For instance, higher education levels correlate with greater income inequality (0.708), likely because those with higher education tend to earn more. A strong negative correlation exists between population density and the elderly population index (−0.923), indicating that densely populated areas tend to have fewer elderly residents.

Overall, the Pearson coefficients indicate no significant multicollinearity (above ±0.8), except for the relationship between population density and the elderly population. To further validate these findings, additional tests, such as Variance Inflation Factor (VIF) analysis, were conducted to ensure that multicollinearity does not distort the results.

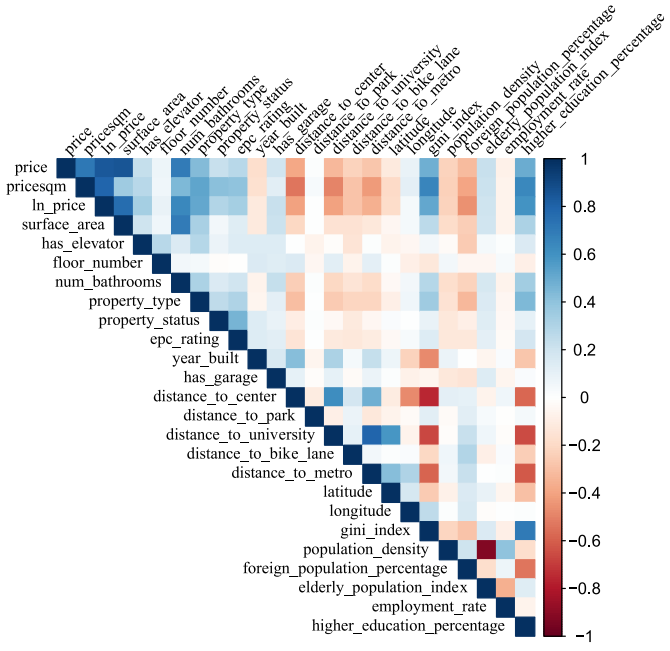


Fig. 3. Correlation heatmap based on Pearson index.

4.5. Segmentation of the real estate market

To effectively segment the Turin apartment market, the hierarchical clustering algorithm was then applied, using a combination of intrinsic property characteristics, such as physical attributes and energy performance, along with accessibility and neighborhood features, including distances to key amenities, and neighborhood dummies to account for spatial features not explicitly included. The study builds on Kim and Irakoze (2022), Hu et al. (2022) and Shi et al. (2015), who focused on clustering based on intrinsic and accessibility characteristics.

The elbow method suggested eight clusters (k = 8) as optimal. The resulting market segmentation revealed distinct sub-markets with similar property value patterns, allowing for a detailed spatial analysis of Turin's real estate landscape (Fig. 4).

Cluster 1 (CL1), including neighborhoods like San Salvario, Cenisia, and Crocetta, has high property values, with a mean price per square meter of €2400. Properties in this cluster tend to have larger surface areas, contributing to their elevated market value and typically good EPC ratings. These areas are desirable due to their proximity to the city center, good public transport, and a mix of historical and modern buildings, attracting a middle-to high-income demographic. Cluster 2 (CL2), covering San Donato, Aurora, Vanchiglia, and Vanchiglietta, has relatively high property values, with a mean price per square meter of €1884. Properties here have moderate surface areas and generally favorable EPC ratings, benefiting from good accessibility, educational facilities, and green spaces. Cluster 3 (CL3), comprising peripheral neighborhoods like Le Vallette, Madonna di Campagna, and Borgata Vittoria, has lower property values, with a mean price per square meter of €1358. Properties here have smaller surface areas and average or below-average EPC ratings, offering affordable housing options.

Cluster 4 (CL4), consisting of Parella and Pozzo Strada, has a mean property price of €149,346 and a median of €120,000, with larger post-war apartments. Its suburban nature, ample green spaces, and family-friendly environment attract middle-income families. Cluster 5 (CL5), including Barriera di Milano, Falchera, and Regio Parco, has the lowest property values, with a mean price per square meter of €1010, appealing to buyers seeking affordable options. Properties generally have lower EPC ratings, reflecting the need for redevelopment and energy efficiency improvements. Cluster 6 (CL6), covering Mirafiori Sud, Mirafiori Santa Rita, and San Paolo, has a mean price per square meter of €1500. It includes both post-war and modern buildings, offering affordable suburban housing with local amenities. It includes both post-war and modern buildings, offering affordable suburban housing with local amenities.

Cluster 7 (CL7), centered around Nizza Millefonti, has a mean price per square meter of €1890. Urban regeneration driven by proximity to the Lingotto complex attracts younger professionals and families. Cluster 8 (CL8), in central Turin, has the highest property values, with a mean price per square meter of €3000. This cluster is distinguished by historical significance, older buildings, and proximity to key services, making it the most exclusive sub-market.

The clustering of Turin's neighborhoods is influenced by property values, building types, accessibility, and socio-demographic factors. Central and semi-central clusters (CL1, CL2, CL8) have higher property values and better amenities, while peripheral clusters (CL3, CL5) offer more affordable housing, appealing to lower-income residents.

4.6. Supervised regression models and linear regression results

The dataset was then divided based on the eight identified clusters, and both overall and cluster-specific analyses were conducted. Given the limited number of cases in each cluster, only training and test sets were utilized. Each dataset was split into training and test sets, with 80% allocated for training and 20% for testing. After the split, all features were standardized to ensure consistency across the dataset. However, to enhance the robustness of the validation process and mitigate

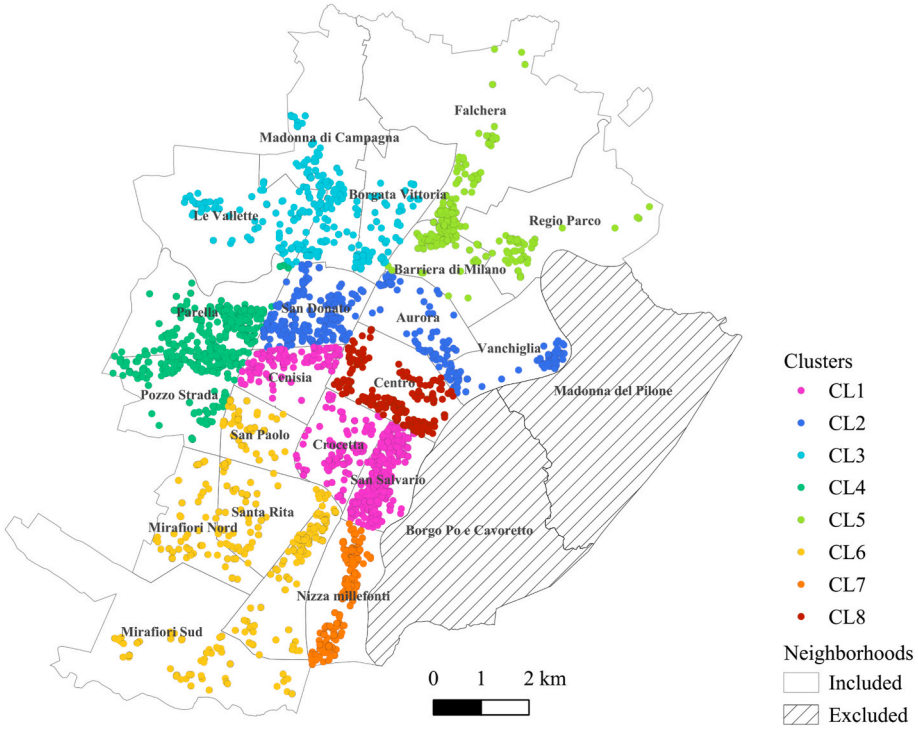


Fig. 4. Clusters identified by hierarchical clustering (k = 8).

overfitting, k-fold cross-validation (k-fold = 10) was applied to the training set.

The evaluation of various ML models applied to the full dataset and individual clusters revealed significant differences in predictive performance (Table 1). For the full dataset, GBM outperformed other models, achieving an R^2 of 0.902, underscoring its effectiveness in capturing complex relationships. RF also demonstrated strong performance, with an R^2 of 0.889 and a relatively low MAE of 0.184.

For individual clusters, GBM consistently delivered the best results, particularly for CL4, where it achieved the lowest RMSE (0.182) and MAE (0.140), along with the highest R^2 (0.910). Conversely, ANN produced mixed results, exhibiting higher error metrics and lower R^2 values, especially for CL7, where it performed the weakest (RMSE of 0.845). The weaker performance of k-NN and ANN compared to

Table 1
Comparison of the performance of LR with ANN, k-NN, SVR, RF, and GBM on test data prediction.

Set	Algorithm	RMSE	MAE	R^2	Algorithm	RMSE	MAE	R^2
All data	LR	0.287	0.217	0.830	SVR	0.283	0.213	0.844
CL1		0.303	0.244	0.819		0.317	0.238	0.800
CL2		0.345	0.254	0.695		0.312	0.232	0.734
CL3		0.306	0.225	0.699		0.267	0.207	0.771
CL4		0.219	0.167	0.875		0.243	0.194	0.839
CL5		0.261	0.203	0.704		0.243	0.185	0.718
CL6		0.193	0.153	0.764		0.235	0.176	0.667
CL7		0.224	0.187	0.819		0.262	0.196	0.783
CL8		0.343	0.251	0.835		0.469	0.330	0.689
All data	ANN	0.264	0.206	0.856	RF	0.237	0.184	0.889
CL1		0.622	0.462	0.601		0.273	0.197	0.852
CL2		0.468	0.362	0.540		0.250	0.202	0.837
CL3		0.607	0.432	0.181		0.231	0.180	0.822
CL4		0.983	0.818	0.497		0.207	0.165	0.883
CL5		0.553	0.462	0.146		**0.222**	**0.179**	**0.765**
CL6		0.267	0.205	0.597		0.189	0.147	0.772
CL7		0.835	0.566	0.070		0.248	0.206	0.782
CL8		0.850	0.621	0.488		**0.265**	**0.215**	**0.905**
All data	k-NN	0.477	0.373	0.558	GBM	**0.217**	**0.166**	**0.902**
CL1		0.519	0.412	0.456		**0.269**	**0.194**	**0.860**
CL2		0.518	0.411	0.406		**0.242**	**0.192**	**0.845**
CL3		0.451	0.370	0.399		**0.227**	**0.177**	**0.832**
CL4		0.487	0.387	0.349		**0.182**	**0.140**	**0.910**
CL5		0.481	0.366	0.111		0.237	0.190	0.740
CL6		0.398	0.303	0.155		**0.161**	**0.126**	**0.836**
CL7		0.416	0.321	0.458		**0.221**	**0.181**	**0.830**
CL8		0.743	0.619	0.294		0.331	0.256	0.849

* Best performance metrics in bold.

ensemble methods like RF and GBM can be attributed to several factors. k-NN struggles with high dimensionality and noise sensitivity, reducing its ability to effectively distinguish between data points in complex datasets. ANN models require extensive hyperparameter tuning and larger datasets to perform well, and they are prone to overfitting, particularly in smaller clusters. In contrast, ensemble methods such as RF and GBM are more robust against overfitting and capture complex non-linear relationships through their iterative learning processes, leading to better performance metrics like lower RMSE and higher R^2 values.

Graphs of actual versus predicted values in Figure A1 further confirm these findings, showing tighter clustering of predicted values around the actual values for GBM and RF, especially in clusters, indicating better model fit and prediction consistency.

4.7. Discussion of the results

Although both GBM and RF exhibited comparable performance, GBM emerged as the preferred model due to its consistent accuracy and robustness across most datasets.

The SHAP analysis of Turin's housing market provides comprehensive insights into the differential impact of various features on property values across neighborhood clusters. It reveals that the significance of features varies substantially between neighborhoods, reflecting the distinct characteristics and dynamics of each area. This variability underscores the complex interplay between building attributes, location, and socio-economic conditions that ultimately shape property values.

Across the entire dataset, surface area, property status, Gini Index, and the proportion of residents with higher education emerged as the primary determinants of property values. Fig. 5a illustrates a clear market preference for energy-efficient properties, with superior EPC ratings generally correlating with higher values. This observation is confirmed by the LR model, where the EPC coefficient of 0.042 suggests a 4.2% increase in property value for each incremental improvement in the EPC rating scale.

Surface area, property status and EPC are the key priorities influencing property values in CL1. Properties in this cluster (Fig. 5b) benefit from proximity to the city center, well-connected public transportation, and a blend of historical and modern buildings, making it attractive to middle- and high-income residents. The LR model identified surface area (coefficient 0.373) and property status (0.115) as significant predictors of property value, with EPC ratings contributing to a 5.5% marginal increase in value due to improved energy efficiency. In CL2, the GBM and LR models identified surface area, property status, and foreign population percentage as significant factors. EPC ratings contribute to a 7% marginal increase in property value according to the LR model. The San Donato neighborhood, adjacent to Parco Dora, has undergone urban regeneration, including the development of new energy-efficient buildings, which has bolstered property values.

Peripheral areas in CL3, such as Le Vallette, Madonna di Campagna, and Borgata Vittoria, are characterized by accessible housing. SHAP analysis indicates that, following surface area, the year of construction is the second most important variable, highlighting a clear preference for newer buildings. Moreover, well-maintained or newly constructed apartments are highly valued. The LR model confirmed surface area (0.296) and property status (0.142) as the most influential factors, while the EPC rating was not significant.

CL4, which includes Parella and Pozzo Strada, is characterized by larger post-war apartments. The suburban character, ample green spaces, and family-friendly environment make these areas appealing to middle-income families. In addition to surface area and property status, the presence of elevators and proximity to metro stations were significant factors, particularly for properties within a small buffer zone around the metro line, which positively influenced property values.

Areas in CL5, adjacent to CL3, including Barriera di Milano, Falchera, and Regio Parco, are characterized by lower property values. Properties typically have smaller surface areas and lower EPC ratings, reflecting the need for redevelopment and energy efficiency improvements. Both GBM and LR models indicated that EPC ratings were not significant in this cluster.

Mirafiori Sud, Mirafiori Santa Rita, and San Paolo (CL6) encompass both post-war and modern buildings, offering affordable suburban housing with access to local amenities. Surface area, distance to the city center, and property status were identified as the primary factors influencing property values. SHAP analysis underscored the importance of EPC ratings in this cluster, supported by the LR model with a coefficient of 0.045.

Property values in CL7, centered around Nizza Millefonti, are influenced by urban regeneration initiatives linked to the Lingotto complex, which attract younger professionals and families. Surface area was the most influential factor, followed by floor number and distance to the city center, while EPC ratings were of limited significance.

CL8, located in central Turin, stands out due to its historical significance, older building stock, and proximity to key services, making it the most exclusive sub-market in the city. Surface area, property status, and distance to the city center were identified as dominant factors influencing property values (Fig. 5i).

The GBM model results, analyzed using SHAP values, provided a nuanced understanding of the factors influencing property values, revealing non-linear relationships and highlighting the significance of EPC ratings across multiple clusters. In contrast, the LR model offered a more straightforward but less comprehensive interpretation of these relationships. Both LR and GBM models identified EPC ratings as significant in clusters CL1, CL2, CL5, CL6, and CL7. The non-linearity captured by SHAP analysis indicated distinct groupings within these clusters, suggesting that the impact of factors such as energy efficiency can vary considerably depending on neighborhood characteristics. The combined use of GBM and SHAP thus provides deeper insights into the complex dynamics at play, effectively modeling real-world scenarios where traditional linear approaches may be insufficient.

Using predictions from the GBM algorithm, PDPs were generated for property attributes, focusing on energy efficiency, property type, and property status, due to their correlation with energy performance. Property type is associated with energy efficiency, as luxury or modern buildings often include advanced energy-saving technologies, whereas economical or older properties generally lack these features. Property status, such as being new or recently renovated, directly affects energy performance, since newer or upgraded buildings tend to incorporate modern energy-efficient systems.

CL1 exhibits a pronounced preference for higher energy performance classes ($\geq$A and B), as shown in Fig. 6. The PDP value decreases markedly from classes A-B to G, underscoring the significance of energy efficiency in determining property values. CL3 displays an upward trend in PDP as energy performance improves, indicating that energy efficiency has a positive impact on property value in this cluster. The observed 'green premium' for class A highlights a considerable increase in valuation for properties with the highest energy efficiency, surpassing other classes. This trend suggests a particular sensitivity to energy performance within this segment of the market.

CL6 reveals a significant discrepancy between properties in groups 'ABCD' and those in groups 'EFG'. This differentiation implies that higher energy performance results in a notable increase in property value, whereas lower-performing classes exhibit minimal variation in valuation among themselves. Thus, energy efficiency exerts a considerable influence on properties with superior ratings, but once the classification falls below a certain threshold, the distinctions between lower classes become less relevant.

In CL7, certain EPC classes are absent. This absence may reflect the characteristics of the housing stock in this area, where properties with exceptionally high energy ratings are rare. The lack of representation of some classes suggests that energy performance is not a predominant factor for this cluster, possibly due to the uniformity in building types or

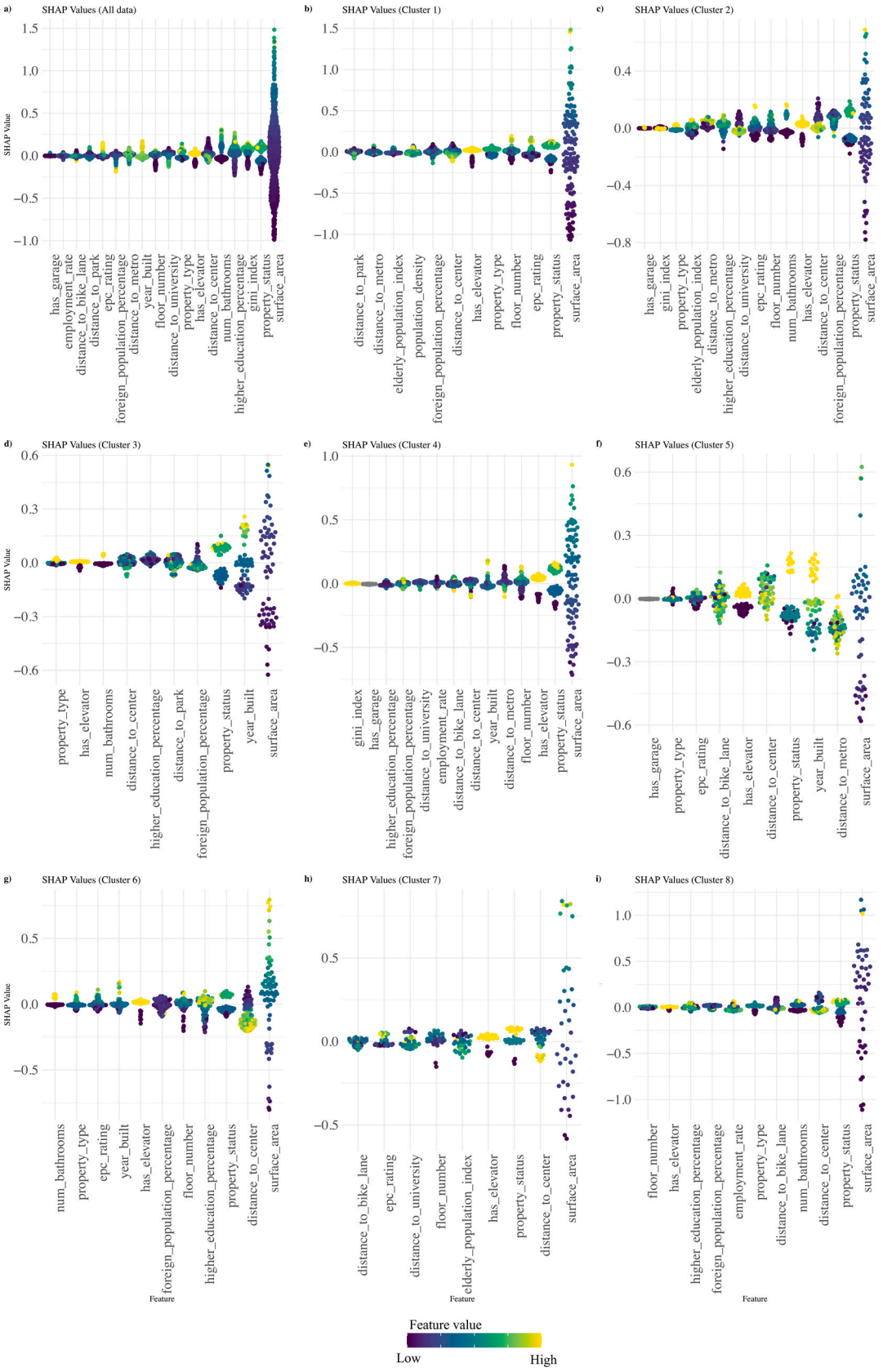


Fig. 5. Beeswarm Plots of SHapley Additive exPlanations (SHAP) values.

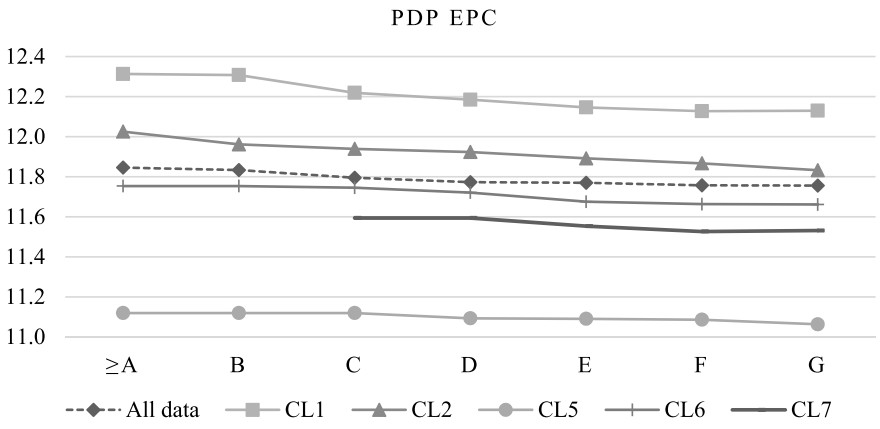


Fig. 6. Partial Dependence Plot (PDP) for EPC estimated by GBM models.

the nature of the housing market in this area.

In contrast, CL5 characterized by affordable housing where priorities other than energy efficiency are emphasized, show limited sensitivity to energy performance. Similarly, energy performance is not a major determinant of property value in CL8. This cluster encompasses the city center, where factors such as location, accessibility, and infrastructure exert a more substantial influence on buyer preferences compared to energy performance.

Fig. 7's PDPs show how property classifications, from economical to luxury, affect predicted values across clusters. Generally, property values rise with higher classifications.

In CL1, there is a notable gap between economical/average and high/luxury properties. Values remain steady from economical to average, but increase sharply for high-class properties, leveling off at the luxury level, indicating a market ceiling where further exclusivity does not significantly boost value.

In CL2 and CL3, the effect of classification is less pronounced, while in CL5, property type has little impact, likely due to the homogeneity of the housing stock.

In the city center (CL8), economical housing is absent, reflecting its prestigious, high-value nature. Similarly, in CL6, luxury properties are rare and have minimal influence on overall values.

Fig. 8 shows the influence of property condition on predicted values across clusters, highlighting a clear preference for well-maintained properties. Most clusters exhibit an upward trend, with better conditions correlating with higher valuations, emphasizing the importance of property maintenance in increasing value. In CL7, the impact on asking prices is consistent across condition levels. However, in CL3 and CL5, where architectural quality is lower, properties in good condition see a more significant value increase. Renovated and new properties are notably valued higher, while those needing renovation consistently show lower valuations.

The study provides several policy recommendations for Turin's residential sector to foster energy efficiency across different urban segments. In clusters like CL3 and CL5, where EPC ratings have minimal impact on property values, affordability is key. Renovation programs should enhance both quality and energy performance, especially in neighborhoods with aging infrastructure. Developing energy communities and positive energy districts could foster economies of scale, making energy-efficient investments more feasible and affordable while improving neighborhood value (Blečić et al., 2023). In semi-central clusters like CL2 and CL6, where EPC ratings significantly impact property values, subsidies for energy efficiency upgrades, especially for older buildings, could boost property values and align with sustainability goals, as residents are willing to invest in greener solutions. In central areas like CL4, EPC ratings, building age, and status influence property values. Due to limited space, policy initiatives should provide incentives for new, energy-efficient construction and retrofitting of historical properties, ensuring modern technologies are integrated while preserving heritage. Incentives could be in the form of tax breaks or grants for sustainable construction.

Beyond localized efforts, revising EPC classifications is essential to make high-efficiency investments more attractive. Highlighting economic benefits, such as cost savings and incentives, and incorporating

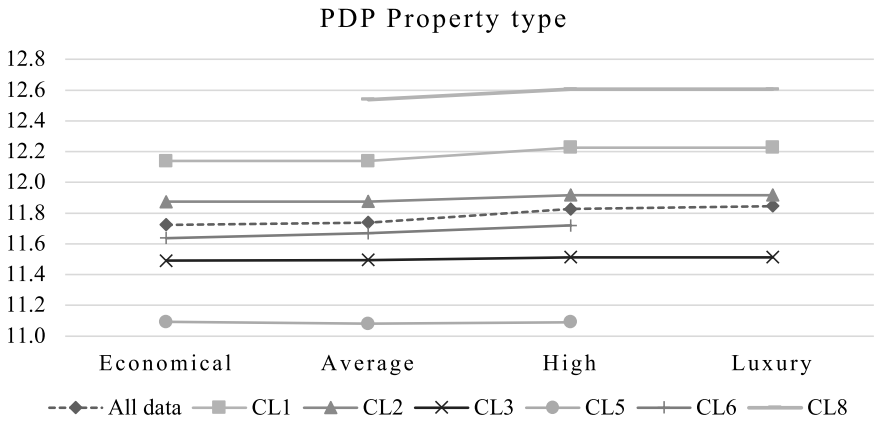


Fig. 7. Partial Dependence Plot (PDP) for property type estimated by GBM models.

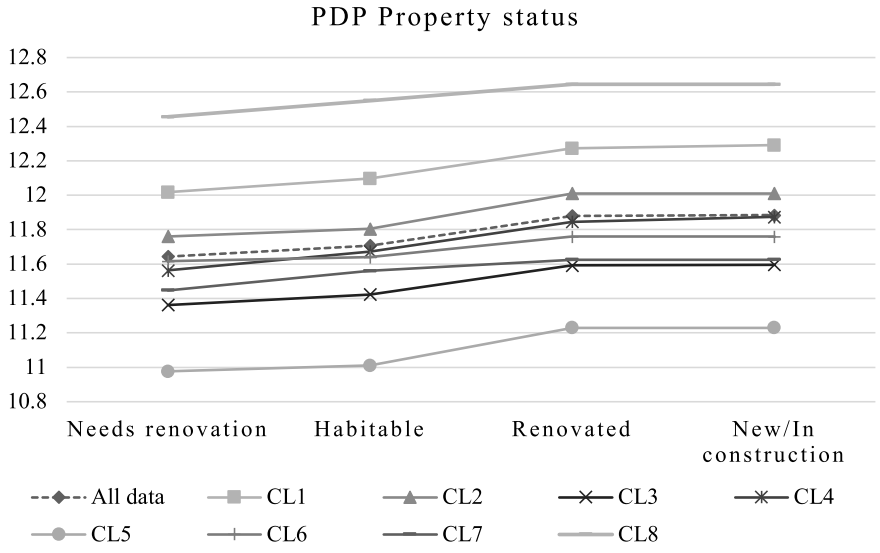


Fig. 8. Partial Dependence Plot (PDP) for property status estimated by GBM models.

additional indicators like carbon emissions and renewable energy use, could align EPC with 2024 EPBD recast requirements, offering a comprehensive view of property sustainability (Junkel, 2022).

5. Conclusion and policy implications

This study provides a comprehensive analysis of how EPC influence property values, using Turin (Northern Italy) as a case study. The standardized framework enhances prediction accuracy and offers detailed insights into the real estate market and consumer behavior. It provides policymakers with valuable tools for assessing EPC impacts and promoting energy efficiency more effectively.

Through hierarchical clustering, the study segments the Turin real estate market into eight clusters based on property attributes and location, offering deeper market insights. By integrating the HPM with advanced ML techniques, the study reveals distinct valuation patterns across segments, enabling more targeted interventions.

Comparing HPM with ML algorithms like GBM, the research shows the superior accuracy of ML in modeling complex relationships between EPC ratings and property values. ML captures non-linear relationships more effectively, providing a precise representation of market dynamics. However, ML 'black box' nature presents challenges (Valier and Micelli, 2020). XAI techniques, such as SHAP values and PDPs, provide transparency and visual insights, making the findings accessible and enabling informed decision-making for stakeholders.

5.1. Policy implications

The ML framework supports tailored policy approaches at the national level while ensuring alignment with EU-wide climate and energy goals. It helps MSs understand how EPC ratings influence property values within their markets, offering valuable insights for designing incentives that encourage energy-efficient home adoption. In markets where EPC ratings are not fully reflected in property valuations, the model can guide governments to introduce subsidies or tax incentives to promote energy-efficient transactions. It can also encourage green financing options like green mortgages, as outlined in initiatives such as the Energy-efficient Mortgages Action Plan (Dell'Anna et al., 2022; EeMAP, 2021), or support broader energy efficiency measures at a large scale, such as the development of energy communities (Barbaro and Napoli, 2024). These actions aim to increase awareness and market recognition of energy-efficient properties and drive investment in energy-efficient solutions.

A key strength of the ML framework is its ability to address regional variations in housing markets across the EU. While EPC are uniformly required under the EPBD, their impact on property values varies based on local factors like energy costs, buyer awareness, climate, and building standards (Bisello et al., 2020; Lyons et al., 2013). For instance, in southern European countries with high energy costs, properties with better EPC ratings often carry a more significant green premium due to potential energy bill savings (Marmolejo-Duarte and Chen, 2019; Taltavull et al., 2017). In contrast, in northern European countries, where sustainability awareness is stronger and building standards are stricter, demand for energy-efficient homes is driven more by environmental considerations and long-term benefits rather than immediate cost savings (Hyland et al., 2013; Jensen et al., 2016).

The standardization of EPC ratings across the EU, mandated by the EPBD, creates a consistent data foundation for the ML framework, enabling its adaptability to various European contexts (European Commission, 2010). Real estate advertisements play a key role in this framework across EU countries. Platforms such as Immobiliare.it and Idealista in Italy, Rightmove and Zoopla in the UK, SeLoger in France, ImmobilienScout24 in Germany, Idealista and Fotocasa in Spain, and Funda in the Netherlands provide essential information, such as property size, location, and energy performance. However, as discussed later, the reliance on asking prices in these listings may limit the accuracy of market analyses.

5.2. Limitations and future perspectives

A key limitation of this study is its generalizability, as the findings from Turin's real estate market may not directly apply to areas with different market dynamics. Nevertheless, the research offers valuable insights into the influence of EPC ratings on property values, with broader implications for the European context. Although the methodology is specific to Turin, it serves as a blueprint for similar analyses elsewhere, demonstrating adaptability to diverse market conditions. The complexity of modeling property characteristics and sustainability features also underscores the need for further research using advanced ML and DL techniques. As sustainability grows in importance to consumers, understanding the evolving value of energy-efficient properties in various settings is crucial.

Another limitation arises from using asking prices rather than transaction prices. Asking prices reflect seller expectations, which may

lead to overestimation of property values, thus introducing bias in assessing the impact of EPC ratings. For instance, asking prices could exaggerate the 'green premium' if sellers anticipate higher demand for energy-efficient homes, whereas transaction prices might reveal a more moderate effect. However, this study primarily focuses on relative trends and differences across clusters, maintaining the validity of the comparative analysis. Future research would benefit from access to transaction-level data to refine these findings.

Finally, although hierarchical clustering provides a robust method for market segmentation, significant potential lies in integrating GWR with ML techniques to address both spatial autocorrelation and market segmentation. Combining the localized insights of GWR with the pattern recognition abilities of ML could yield more refined models that capture the complex relationships between property values and their spatial determinants. This integration would allow for the simultaneous treatment of spatial heterogeneity and autocorrelation while leveraging ML's predictive power. For example, after applying GWR to capture localized effects, the resulting spatially varying coefficients could be used as input for clustering algorithms like k-means, enabling more nuanced segmentation based on both spatial and non-spatial factors (Hu et al., 2022; Marmolejo-Duarte et al., 2020).

This framework could significantly improve the accuracy and reliability of the assessment of the effectiveness of EPC in different contexts, providing a holistic view of real estate dynamics in Europe. However, future research could explore these integrations to better address the complexity of spatially dependent real estate data and provide policymakers with insights.

Declaration of competing interest

The author declares the following financial interests/personal relationships which may be considered as potential competing interests: Federico Dell'Anna reports administrative support, article publishing charges, equipment, and writing assistance were provided by Politecnico di Torino.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.enpol.2024.114407.

Data availability

The author does not have permission to share data.

References

Akhtyrska, Y., Fuerst, F., 2024. The effectiveness of climate change regulations in the commercial real estate market. Energy Pol 185, 113916. https://doi.org/10.1016/j.enpol.2023.113916.

Alonso, W., 1964. Location and Land Use. Harvard University Press, Cambridge. https://doi.org/10.4159/harvard.9780674730854.

Al-Qawasmi, J., 2022. Machine learning applications in real estate: critical review of recent development. In: IFIP Advances in Information and Communication Technology. Springer Science and Business Media Deutschland GmbH, pp. 231–249. https://doi.org/10.1007/978-3-031-08337-2_20.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. Am. Statistician 46, 175. https://doi.org/10.2307/2685209.

Ara Aksoy, S., Irwin, E., 2021. Cluster analysis for housing market segmentation. Sosyoekonomi 29, 11–32. https://doi.org/10.17233/sosyoekonomi.2021.03.01.

Aydin, Y.C., Mirzaei, P.A., Akhavannasab, S., 2019. On the relationship between building energy efficiency, aesthetic features and marketability: toward a novel policy for energy demand reduction. Energy Pol. 128, 593–606. https://doi.org/10.1016/J.ENPOL.2018.12.036.

Banfi, S., Farsi, M., Filippini, M., Jakob, M., 2008. Willingness to pay for energy-saving measures in residential buildings. Energy Econ. 30, 503–516. https://doi.org/10.1016/j.eneco.2006.06.001.

Barbaro, S., Napoli, G., 2024. Towards a participatory energy transition: Critical issues and potentials of regulatory and financial instruments for Renewable Energy Communities (RECs) in Italy. Valori e Valutazioni 35, 69–95. https://doi.org/10.48264/VVSIEV-20243506.

Barreca, A., Fregonara, E., Rolando, D., 2021. EPC labels and building features: spatial implications over housing prices. Sustainability 13, 2838. https://doi.org/10.3390/su13052838.

Becchio, C., Bottero, M.C., Corgnati, S.P., Dell'Anna, F., 2018. Evaluating health benefits of urban energy retrofitting: An application for the city of Turin. In: Bisello, A., Vettorato, D., Laconte, P., Costa, S. (Eds.), Smart and Sustainable Planning for Cities and Regions. SSPCR 2017. Green Energy and Technology. Springer, Cham, pp. 281–304. https://doi.org/10.1007/978-3-319-75774-2_20.

Bergstra, J., Ca, J.B., Ca, Y.B., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.

Berto, R., Tintinaglia, F., Rosato, P., 2023. How much is the indoor comfort of a residential building worth? A discrete choice experiment. Build. Environ. 245, 110911. https://doi.org/10.1016/j.buildenv.2023.110911.

Bhagat, H., Priya, S., Aditya, K., 2020. Outlier detection based on machine learning techniques. Int. J. Adv. Sci. Technol. 29, 2142–2151.

Birant, D., Kut, A., 2007. ST-DBSCAN: an algorithm for clustering spatial–temporal data. Data Knowl. Eng. 60, 208–221. https://doi.org/10.1016/j.datak.2006.01.013.

Bisello, A., Antoniucci, V., Marella, G., 2020. Measuring the price premium of energy efficiency: a two-step analysis in the Italian housing market. Energy Build. 208, 109670. https://doi.org/10.1016/j.enbuild.2019.109670.

Blečić, I., Carrus, A.S., Muroni, E., Saiu, V., Saliu, M.C., 2023. Engagement and Inclusion Experiences for Energy Communities: An Ongoing Case Study in Cagliari, Italy. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 14109 LNCS,"first and last page" 513–528. In: 23rd International Conference on Computational Science and Its Applications, ICCSA 2023. Springer, Athens. https://doi.org/10.1007/978-3-031-37120-2_33, 3–6 July 2023.

Bottero, M., Bravi, M., Cavana, G., Dell'Anna, F., Becchio, C., Corgnati, S.P., 2019. Retrofit energetico e decisioni di investimento: valutazione delle preferenze degli individui attraverso un esperimento di scelta (Energy retrofit and investment decisions: individuals' preferences valuation through a Choice Experiment). Geoingegner. Ambientale e Miner. 158, 11–24.

Bottero, M., Bravi, M., Dell'Anna, F., Mondini, G., 2018. Valuing building energy efficiency through Hedonic Prices Method: are spatial effects relevant? Valori e Valutazioni 21, 27–39.

Bottero, M., Caprioli, C., Foth, M., Mitchell, P., Rittenbruch, M., Santangelo, M., 2022. Urban parks, value uplift, and green gentrification: An application of the spatial hedonic model in the city of Brisbane. Urban For. Urban Green. 74, 127618. https://doi.org/10.1016/j.ufug.2022.127618.

Bourassa, S.C., Hamelink, F., Hoesli, M., Macgregor, B.D., 1999. Defining housing submarkets. J. Hous. Econ. 8, 160–183. https://doi.org/10.1006/jhec.1999.0246.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Brown, D., Sorrell, S., Kivimaa, P., 2019. Worth the risk? An evaluation of alternative finance mechanisms for residential retrofit. Energy Pol. 128, 418–430. https://doi.org/10.1016/J.ENPOL.2018.12.033.

Buso, T., Dell'Anna, F., Becchio, C., Bottero, M.C., Corgnati, S.P., 2017. Of comfort and cost: Examining indoor comfort conditions and guests' valuations in Italian hotel rooms. Energy Res. Social Sci. 32, 94–111. https://doi.org/10.1016/j.erss.2017.01.006.

Cajias, M., 2021. Artificial intelligence and real estate - not just an evolution, a real game changer. J. Property Invest. Finance 39, 15–18. https://doi.org/10.1108/JPIF-06-2020-0063.

Cajias, M., Fuerst, F., Bienert, S., 2019. Tearing down the information barrier: the price impacts of energy efficiency ratings for buildings in the German rental market. Energy Res. Social Sci. 47, 177–191. https://doi.org/10.1016/j.erss.2018.08.014.

Can, A., 1992. Specification and estimation of hedonic housing price models. Reg. Sci. Urban Econ. 22, 453–474. https://doi.org/10.1016/0166-0462(92)90039-4.

Cawley, G.C., Talbot, N.L.C., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. 11, 2079–2107.

Chegut, A., Kok, N., Piet, E., 2011. The value of green buildings new evidence from the United Kingdom. ERES 2010, 1–44.

Copiello, S., Donati, E., 2021. Is investing in energy efficiency worth it? Evidence for substantial price premiums but limited profitability in the housing sector. Energy Build. 251, 111371. https://doi.org/10.1016/j.enbuild.2021.111371.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/BF00994018.

Costa, O., Fuerst, F., Robinson, S.J., Mendes-Da-Silva, W., 2018. Green label signals in an emerging real estate market. A case study of Sao Paulo, Brazil. J. Clean. Prod. 184, 660–670. https://doi.org/10.1016/j.jclepro.2018.02.281.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. 13, 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Crespo Sánchez, E., Spairani Berrio, S., Onecha Perez, B., Marmolejo-Duarte, C., 2021. Perceived benefits of energy efficiency in the Spanish residential market and their relation to sociodemographic and living conditions. Appl. Sci. 11, 875. https://doi.org/10.3390/app11020875.

Cropper, M.L., Deck, L.B., McConnell, K.E., 1988. On the choice of funtional form for hedonic price functions. Rev. Econ. Stat. 70, 668. https://doi.org/10.2307/1935831.

Curto, A., Fregonara, E., Semeraro, P., 2012. Asking prices vs. Market prices: an empirical analysis. Territorio 1, 53–72.

D'Alpaos, C., Bragolusi, P., 2018. Buildings energy retrofit valuation approaches: state of the art and future perspectives. Valori e Valutazioni 20, 79–94.

Del Giudice, F.P., Manganelli, B., De Paola, P., Tajani, F., Amato, F., 2024. An Analysis of the Airbnb Market: A Detailed Look at Four Italian Cities. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 14822 LNCS. In: 24th International Conference on

Computational Science and Its Applications, ICCSA 2024. Springer, Hanoi. https://doi.org/10.1007/978-3-031-65318-6_4. (Accessed 14 July 2024).
Dell'Anna, F., Bottero, M., 2021. Green premium in buildings: evidence from the real estate market of Singapore. J. Clean. Prod. 286, 125327. https://doi.org/10.1016/J.JCLEPRO.2020.125327.
Dell'Anna, F., Marmolejo-Duarte, C., Bravi, M., Bottero, M., 2022. A choice experiment for testing the energy-efficiency mortgage as a tool for promoting sustainable finance. Energy Effic. 15 (5), 27. https://doi.org/10.1007/s12053-022-10035-y.
Deng, Y., Li, Z., Quigley, J.M., 2012. Economic returns to energy-efficient investments in the housing market: evidence from Singapore. Reg. Sci. Urban Econ. 42, 506–515. https://doi.org/10.1016/j.regsciurbeco.2011.04.004.
Ding, C., Liu, T., Cao, X., Tian, L., 2022. Illustrating nonlinear effects of built environment attributes on housing renters' transit commuting. Transp. Res. D Transp. Environ. 112, 103503. https://doi.org/10.1016/j.trd.2022.103503.
EeMAP, 2021. Energy-efficient Mortgages Action Plan: Energy Performance Indicators. Available at: https://www.energyefficientmortgages.eu/wp-content/uploads/2021/07/EEMI-Energy-Performance-Indicators.pdf. (Accessed 13 September 2024).
Encinas, F., Marmolejo-Duarte, C., Sánchez de la Flor, F., Aguirre, C., 2018. Does energy efficiency matter to real estate-consumers? Survey evidence on willingness to pay from a cost-optimal analysis in the context of a developing country. Energy Sustain. Dev. 45, 110–123. https://doi.org/10.1016/j.esd.2018.05.008.
European Commission, 2019. Ethics Guidelines for Trustworthy AI. European Commission, Brussels. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. (Accessed 12 October 2024).
Dell'Anna, F. (2022). Spatial Econometric Analysis of Multi-family Housing Prices in Turin: The Heterogeneity of Preferences for Energy Efficiency. In Gervasi, O., Murgante, B., Misra, S., Rocha, A.M.A.C., Garau, C. (Eds.), Computational Science and Its Applications – ICCSA 2022 Workshops. Lecture Notes in Computer Science. vol. 13380, 211–227. Springer, Cham. https://doi.org/10.1007/978-3-031-10542-5_15.
Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR, USA, pp. 226–231.
European Commission, 2024. Questions and Answers on the revision of the Energy Performance of Buildings Directive [WWW Document]. URL. https://ec.europa.eu/commission/presscorner/detail/en/qanda_24_1966#:~:text=56.022%2CkB-PDF-,Download,-Contactsformedia.
European Commission, 2010. Directive 2010/31/UE. Energy Perf. Build. Direct. (EPBD) [WWW Document]. URL. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32010L0031&from=it. (Accessed 2 August 2019).
European Commission, 2002. Directive 2002/91/CE. Energy Perf. Build. Direct. (EPBD) [WWW Document]. URL. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32002L0091&from=EN. (Accessed 2 August 2019).
Ferreira, M., Almeida, M., 2015. Benefits from energy related building renovation beyond costs, energy and emissions. Energy Procedia 78, 2397–2402. https://doi.org/10.1016/j.egypro.2015.11.199.
Fregonara, E., Rolando, D., Semeraro, P., Vella, M., 2014. The impact of Energy Performance Certificate level on house listing prices. First evidence from Italian real estate. Aestimum 65, 143–163. https://doi.org/10.13128/Aestimum-15459.
Fryer, D., Strümke, I., Nguyen, H., 2020. Shapley value confidence intervals for attributing variance explained. Front. Appl. Math Stat. 6. https://doi.org/10.3389/fams.2020.587199.
Fuerst, F., McAllister, P., Nanda, A., Wyatt, P., 2016a. Energy performance ratings and house prices in Wales: an empirical study. Energy Pol. 92, 20–33. https://doi.org/10.1016/j.enpol.2016.01.024.
Fuerst, F., McAllister, P., Nanda, A., Wyatt, P., 2015. Does energy efficiency matter to home-buyers? An investigation of EPC ratings and transaction prices in England. Energy Econ. 48, 145–156. https://doi.org/10.1016/j.eneco.2014.12.012.
Fuerst, F., Oikarinen, E., Harjunen, O., 2016b. Green signalling effects in the market for energy-efficient residential buildings. Appl. Energy 180, 560–571. https://doi.org/10.1016/j.apenergy.2016.07.076.
Gabrielli, L., Giuffrida, S., Trovato, M.R., 2019. Real estate landscapes and the historic city: on how looking inside the market. In: International Journal for Housing Science and its Applications. Springer Science and Business Media Deutschland GmbH, pp. 269–276. https://doi.org/10.1007/978-3-319-92102-0_29.
Galster, G., 1996. William Grigsby and the Analysis of Housing Sub-markets and Filtering. Urban Stud. 33 (10), 1797–1805. https://doi.org/10.1080/0042098966376.
Gao, Q., Shi, V., Pettit, C., Han, H., 2022. Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia. Land Use Pol. 123. https://doi.org/10.1016/j.landusepol.2022.106409.
Goodman, A.C., Thibodeau, T., Goodman, A.C., Thibodeau, T., 2003. Housing market segmentation and hedonic prediction accuracy. J. Hous. Econ. 12, 181–201.
Goodman, A.C., Thibodeau, T.G., 1998. Housing market segmentation. J. Hous. Econ. 7, 121–143. https://doi.org/10.1006/jhec.1998.0229.
Gružauskas, V., Čalnerytė, D., Fyleris, T., Kriščiūnas, A., 2021. Application of multivariate time series cluster analysis to regional socioeconomic indicators of municipalities. Real Estate Manag. Valuat. 29, 39–51. https://doi.org/10.2478/remav-2021-0020.
Heidari, M., Zad, S., Rafatirad, S., 2021. Ensemble of supervised and unsupervised learning models to predict a profitable business decision. In: 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). IEEE, pp. 1–6. https://doi.org/10.1109/IEMTRONICS52119.2021.9422649.
Hodge, V.J., Austin, J., 2004. A survey of outlier detection methodologies. Artif. Intell. Rev. 22 (2), 85–126. https://doi.org/10.1007/S10462-004-4304-Y.
Hu, Lirong, He, S., Su, S., 2022. A novel approach to examining urban housing market segmentation: Comparing the dynamics between sales submarkets and rental submarkets. Comput. Environ. Urban Syst. 94, 101775. https://doi.org/10.1016/j.compenvurbsys.2022.101775.
Hyland, M., Lyons, R.C., Lyons, S., 2013. The value of domestic building energy efficiency - evidence from Ireland. Energy Econ. 40, 943–952. https://doi.org/10.1016/j.eneco.2013.07.020.
Ja'afar, N.S., Mohamad, J., Ismail, S., 2021. Machine learning for property price prediction and price valuation: a systematic literature review. Plan. Malays. 19, 411–422. https://planningmalaysia.org/index.php/pmj/article/download/1018/714/1898.
Jamil, S., Mohd, T., Masrom, S., Ab Rahim, N., 2020. Machine learning price prediction on green building prices. In: 2020 IEEE Symposium on Industrial Electronics & Applications (ISIEA). IEEE, pp. 1–6. https://doi.org/10.1109/ISIEA49364.2020.9188114.
Jensen, O.M., Hansen, A.R., Kragh, J., 2016. Market response to the public display of energy performance rating at property sales. Energy Pol. 93, 229–235. https://doi.org/10.1016/J.ENPOL.2016.02.029.
Junkel, J., 2022. Advancing energy performance certificates to next generation.
Kim, D.H., Irakoze, A., 2022. Identifying market segment for the assessment of a price premium for green certified housing: a cluster analysis approach. Sustainability 15, 507. https://doi.org/10.3390/su15010507.
Konhäuser, K., Werner, T., 2024. Uncovering the financial impact of energy-efficient building characteristics with eXplainable artificial intelligence. Appl. Energy 374, 123960. https://doi.org/10.1016/J.APENERGY.2024.123960.
Lancaster, K.J., 1966. A new approach to consumer theory. J. Polit. Econ. 74, 132–157. https://doi.org/10.1086/259131.
Lance, G.N., Williams, W.T., 1967. A general theory of classificatory sorting strategies: 1. Hierarchical systems. Comput. J. 9, 373–380. https://doi.org/10.1093/COMJNL/9.4.373.
Lee, C., 2021. Data augmentation using a variational autoencoder for estimating property prices. Property Manag. 39, 408–418. https://doi.org/10.1108/PM-09-2020-0057.
Levantesi, S., Piscopo, G., 2020. The importance of economic variables on London real estate market: a random forest approach. Risks 8, 1–17. https://doi.org/10.3390/risks8040112.
Liu, Q., Deng, M., Shi, Y., Wang, J., 2012. A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. Comput. Geosci. 46, 296–309. https://doi.org/10.1016/j.cageo.2011.12.017.
Lundberg, S., 2018. An Introduction to Explainable AI with Shapley Values — SHAP Latest Documentation [WWW Document]. URL. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An/introduction/to/explainable/AI/with/Shapley/values.html. (Accessed 15 October 2024).
Lundberg, S., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA, USA.
Lyons, R., Cohen, F., Lyons, R., Fedrigo-Fazio, D., 2013. Energy performance certificates in buildings and their impact on transaction prices and rents in selected EU countries. Final Rep. Prep. Europ. Comm., DG Energy [WWW Document]. URL. https://ec.europa.eu/energy/sites/ener/files/documents/20130619-energy_performance_certificates_in_buildings.pdf. (Accessed 2 August 2019).
Maclennan, D., Tu, Y., 1996. Economic perspectives on the structure of local housing systems. Hous. Stud. 11, 387–406. https://doi.org/10.1080/02673039608720864.
Manganelli, B., Pontrandolfi, P., Azzato, A., Murgante, B., 2014. Using geographically weighted regression for housing market segmentation. Int. J. Bus. Intell. Data Min. 9, 161–177. https://doi.org/10.1504/IJBIDM.2014.065100.
Marmolejo Duarte, C., 2016. La incidencia de la calificación energética sobre los valores residenciales: un análisis para el mercado plurifamiliar en Barcelona. Inf. Construcción 68, e156. https://doi.org/10.3989/ic.16.053.
Marmolejo-Duarte, C., Chen, A., 2022a. Uncovering the price effect of energy performance certificate ratings when controlling for residential quality. Renew. Sustain. Energy Rev. 166, 112662. https://doi.org/10.1016/j.rser.2022.112662.
Marmolejo-Duarte, C., Chen, A., 2022b. The effect of energy performance ratings over residential prices or how an insufficient control of architectural-quality may render spurious conclusions. Cities 126, 103674. https://doi.org/10.1016/j.cities.2022.103674.
Marmolejo-Duarte, C., Chen, A., 2019. The uneven price impact of energy efficiency ratings on housing segments and implications for public policy and private markets. Sustainability 11, 372. https://doi.org/10.3390/su11020372.
Marmolejo-Duarte, C., Chen, A., Bravi, M., 2020. Spatial implications of EPC rankings over residential prices. In: Mondini, G., Oppio, A., Stanghellini, S., Bottero, M.C., Abastante, F. (Eds.), Values and Functions for Future Cities. Springer, Cham, pp. 51–71. https://doi.org/10.1007/978-3-030-23786-8_4.
Masrom, S., Mohd, T., Rahman, A.S.A., 2022. Green building factor in machine learning based condominium price prediction. IAES Int. J. Artif. Intell. 11, 291–299. https://doi.org/10.11591/ijai.v11.i1.pp291-299.
Mazzitelli, F., Moine, B., 2022. Statistiche regionali Il mercato immobiliare residenziale. Turin.
McCluskey, W.J., Borst, R.A., 2011. Detecting and validating residential housing submarkets. Int. J. Hous. Mark. Anal. 4, 290–318. https://doi.org/10.1108/17538271111153040.
McCord, M., Lo, D., Davis, P.T., Hemphill, L., McCord, J., Haran, M., 2020. A spatial analysis of EPCs in the Belfast Metropolitan Area housing market. J. Property Res. 37, 25–61. https://doi.org/10.1080/09599916.2019.1697345.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133. https://doi.org/10.1007/BF02478259.
Mete, M.O., Yomralioglu, T., 2023. A hybrid approach for mass valuation of residential properties through geographic information systems and machine learning integration. Geogr. Anal. 55, 535–559. https://doi.org/10.1111/gean.12350.
Micelli, E., Giliberto, G., Righetto, E., Tafuri, G., 2023. The economic value of sustainability. Real estate market and energy performance of homes. Valori e Valutazioni 34, 3–16. https://doi.org/10.48264/VVSIEV-20233402.
Michelangeli, A., 2008. l metodo dei prezzi edonici per la costruzione di indici dei prezzi per il mercato immobiliare. In: Del Giudice, V., D'Amato, M. (Eds.), Principi Metodologici Per La Costruzione Di Indici Dei Prezzi Nel Mercato. Maggioli Editore, Dogana, pp. 102–113.
Mohd, T., Jamil, S., Masrom, S., Rahim, N.A., 2022. Machine learning predictive model for green building price. Malays. Construct. Res. J. 16, 156–165.
Murtagh, F., Contreras, P., 2012. Algorithms for hierarchical clustering: an overview. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2, 86–97. https://doi.org/10.1002/WIDM.53.
Napoli, G., Giuffrida, S., Valenti, A., 2017. Forms and Functions of the Real Estate Market of Palermo (Italy). Science and Knowledge in the Cluster Analysis Approach. In: Stanghellini, S., Morano, P., Bottero, M., Oppio, A. (Eds.), Appraisal: From Theory to Practice. Green Energy and Technology. Springer, Cham, pp. 191–202. https://doi.org/10.1007/978-3-319-49676-4_14.
Owen, A., Middlemiss, L., Brown, D., Davis, M., Hall, S., Bookbinder, R., Brisbois, M.C., Cairns, I., Hannon, M., Mininni, G., 2023. Who applies for energy grants? Energy Res. Social Sci. 101, 103123. https://doi.org/10.1016/j.erss.2023.103123.
Palm, R., 1978. Spatial segmentation of the urban housing market. Econ. Geogr. 54, 210. https://doi.org/10.2307/142835.
Pearson, K., 1896. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. JSTOR 187, 253–318.
Raushan, K., Mac Uidhir, T., Llorens Salvador, M., Norton, B., Ahern, C., 2024. A data-driven standardised generalisable methodology to validate a large energy performance Certification dataset: a case of the application in Ireland. Energy Build. 323, 114774. https://doi.org/10.1016/j.enbuild.2024.114774.
Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. J. Polit. Econ. 82, 34–55. https://doi.org/10.1086/260169.
Schnare, A.B., Struyk, R.J., 1976. Segmentation in urban housing markets. J. Urban Econ. 3 (2), 146–166. https://doi.org/10.1016/0094-1190(76)90050-4.
Schuitema, G., Aravena, C., Denny, E., 2020. The psychology of energy efficiency labels: trust, involvement, and attitudes towards energy performance certificates in Ireland. Energy Res. Social Sci. 59, 101301. https://doi.org/10.1016/j.erss.2019.101301.
Shi, C., Wei, B., Wei, S., et al., 2021. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. J. Wirel. Commun. Netw. 2021, 31. https://doi.org/10.1186/s13638-021-01910-w.
Shi, D., Guan, J., Zurada, J., Levitan, A.S., 2015. An innovative clustering approach to market segmentation for improved price prediction. J. Int. Technol. Inform. Manag. 24. https://doi.org/10.58729/1941-6679.1033.
Skovajsa, Š., 2023. Review of clustering methods used in data-driven housing market segmentation. Real Estate Manag. Valuat. 31, 67–74. https://doi.org/10.2478/remav-2023-0022.
Taltavull de La Paz, P., Perez-Sanchez, V., Mora-Garcia, R.T., Perez-Sanchez, J.-C., 2019. Green premium evidence from climatic areas: a case in southern Europe, alicante (Spain). Sustainability 11, 686. https://doi.org/10.3390/su11030686.
Taltavull, P., Anghel, I., Ciora, C., 2017. Impact of energy performance on transaction prices. J. Europ. Real Est. Res. 10, 57–72. https://doi.org/10.1108/JERER-12-2016-0046.
Thackway, W.T., Ng, M.K.M., Lee, C.L., Shi, V., Pettit, C.J., 2022. Spatial variability of the 'Airbnb effect': A spatially explicit analysis of Airbnb's impact on housing prices in Sydney. ISPRS Int. J. Geo-Inf. 11 (1), 65. https://doi.org/10.3390/ijgi11010065.
Tsai, I.C., 2022. Value capitalization effects of green buildings: a new insight through time trends and differences in various price levels. Build. Environ. 224, 109577. https://doi.org/10.1016/J.BUILDENV.2022.109577.
Ürge-Vorsatz, D., Herrero, S.T., Dubash, N.K., Lecocq, F., 2014. Measuring the Co-benefits of climate change mitigation. Annu. Rev. Environ. Resour. 39, 549–582. https://doi.org/10.1146/annurev-environ-031312-125456.
Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J., 2019. Machine learning algorithm validation with a limited sample size. PLoS One 14, e0224365. https://doi.org/10.1371/journal.pone.0224365.
Valier, A., Micelli, E., 2020. Automated models for value prediction: a critical review of the debate. Valori e Valutazioni 24, 151–161.
Wang, X., Lu, M., Mao, W., Ouyang, J., Zhou, B., Yang, Y., 2015. Improving benefit-cost analysis to overcome financing difficulties in promoting energy-efficient renovation of existing residential buildings in China. Appl. Energy 141, 119–130. https://doi.org/10.1016/j.apenergy.2014.12.001.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244. https://doi.org/10.1080/01621459.1963.10500845.
Wiersma, S., Just, T., Heinrich, M., 2022. Segmenting German housing markets using principal component and cluster analyses. Int. J. Hous. Mark. Anal. 15, 548–578. https://doi.org/10.1108/IJHMA-01-2021-0006.
Yağmur, A., Kayakuş, M., Terzioğlu, M., 2023. House price prediction modeling using machine learning techniques: a comparative study. Aestimum 81, 39–51. https://doi.org/10.36253/aestim-13703.
Zhao, Y., Chetty, G., Tran, D., 2023. Real estate price prediction on GenerativeLanguage models. In: 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). IEEE, pp. 1–7. https://doi.org/10.1109/CSDE59766.2023.10487658.