Centre for Research on Settlements and Urbanism

Journal of Settlements and Spatial Planning

Journal homepage: https://jssp.reviste.ubbcluj.ro

The Effect of Dimensionality Reduction on the Real Estate Appraisal Performance Using Tree-Based Machine Learning Models

Jamal A. A. NUMAN[1], Izham Mohamad YUSOFF*[2]

** Corresponding author*

[1] Universiti Sains Malaysia, Institute of Postgraduate Studies, Penang, MALAYSIA

[2] Universiti Sains Malaysia, Geography Section, Transdisciplinary Research on Environmental Science, Occupational Safety and Health, Penang, MALAYSIA

✉ jamalnuman@student.usm.my 0009-0003-7861-4790

✉ izham@usm.my 0000-0003-0805-804X

DOI: https://doi.org/10.24193/JSSP.2025.1.02
Received: 03 August 2024
Received in revised form: 22 January 2025
Accepted for publication: 03 June 2025
Available online: 20 June 2025

Keywords: dimensionality reduction, feature selection, feature importance, Tree-Based models, linear regression

ABSTRACT

Real estate appraisal is a critical process essential for economic, financial, and business transactions, including buying and selling, mortgage lending, insurance, and property taxation. In this context, model-based real estate appraisal methods face significant challenges such as performance, interpretability, stability, reliability, scalability, flexibility, simplicity, adaptability, applicability, generalizability, comprehensibility, data availability, and evaluation metrics. Among these challenges, performance consistently stands out as a key concern, attracting considerable attention from both academic researchers and industry professionals. With the aim of investigating the effect of dimensionality reduction (DR) on the appraisal performance, three objectives are crafted: identifying the initial features affecting real estate appraisal within Al Bireh city, Palestine, selecting the most influential features, and evaluating model performance when all features are included versus when only the most influential are used employing five statistical metrics. The originality of this research lies in the explicit implementation of DR using multiple feature importance (FI) techniques, multiple models, and multiple evaluation metrics. Specifically, this study includes two FI techniques—namely, inherent FI and Shapley Additive Explanation (SHAP); four models - three tree-based models (decision tree (DT), random forest (RF), and extreme gradient boosting (XGBoost)) and a linear regression (LR) model used as a benchmark; and five evaluation metrics: MSE, RMSE, MAE, MAPE, and R². The results indicate no performance improvement when DR is conducted. However, with DR reducing the features from 28 to 6, the relative performance metric decrease is minor, remaining below 5% for all models except LR, and as low as 0.7% in terms of R² for RF, thus concluding the need for a trade-off between the minor decrease in performance and gains in computational efficiency, hardware resources, and data collection. The key implications of DR provide stakeholders with a checklist of key features influencing appraisal value, and increase efficiency by reducing processing time, resources, and data collection.

1. INTRODUCTION

Real estate appraisal, also known as property valuation or house pricing, plays a vital role as a decision-support tool in various significant economic, financial, and business transactions. These include property sales and purchases, bank loans, insurance assessments, property taxation, ownership transfers, partnership terminations, expropriations, settlements, auctions, and other associated activities (Droj et al., 2024; Jin et al., 2024; Oust et al., 2023; Alzain et al., 2022; Mankad, 2022; Sisman and Aydinoglu, 2022; Steurer et al., 2021; Xu and Zhang, 2021). In this regard, model-based real estate appraisal methods encounter a myriad of challenges, including issues related to interpretability, stability, reliability, scalability, flexibility, simplicity, adaptability, applicability, generalizability, comprehensibility, data availability, and evaluation metrics. Among these, performance has consistently emerged as a prominent concern, drawing significant attention from both academic researchers and industry stakeholders over the years (Chen et al., 2024; Elnaeem Balila and Shabri, 2024; Hoxha, 2024; Hurley and Sweeney, 2024; Jin et al., 2024; Mathotaarachchi et al., 2024; Song and Ma, 2024; Çılgın and Gökçen, 2023; Geerts and De Weerdt, 2023; Hoang and Wiegratz, 2023; Lahmiri et al., 2023; Oust et al., 2023; Stang et al., 2023; Zhan et al., 2023; Das et al., 2021). This stems from the fact that higher model performance leads to more mature and informed decision-making, resulting in higher-quality and more impactful outcomes in practice.

With the aim of measuring the effect of DR on real estate appraisal performance improvement, three questions are articulated in this research: What features influence real estate appraisal within Al Bireh city? How can multiple FI techniques and tree-based machine learning methods be combined to select the most influential features? How can performance be evaluated using the full set of initial features versus the most influential ones? The rationale for implementing DR, which involves eliminating complexity, noise, redundancy, and irrelevance, is not only to reduce computational time, hardware resources such as CPU, memory and disk space, and data collection efforts, but also to enhance performance, explainability, and stability (Mallick and Mittal, 2025; Theng and Bhoyar, 2024; Zouhri et al., 2024; Chhikara et al., 2022; Chanasit et al., 2021; Palo et al., 2021).

In response to the first question, this study determines the initial features affecting real estate appraisal by drawing on a literature review, relevant local laws and regulations in Al Bireh city - the study area - and the authors' professional expertise in the field (Numan and Yusoff, 2024b). Regardless of the dimensionality of the potential features identified, they undergo an analysis process to select the most influential ones, enabling a comparison of model performance using all potential features versus only the most influential ones.

Concerning the second research question, which focuses on combining the results of multiple FI techniques employed within various models to select a single set of the most influential features, some background on the DR method is necessary. Essentially, DR includes two major approaches: feature extraction and feature selection (Jia et al., 2022; Palo et al., 2021). Feature extraction reduces the original number of features by creating a new set, rather than a subset, derived from the original features through methods like Principal Component Analysis (PCA) (Lee, 2021). Feature selection, on the other hand, reduces the number of features by selecting a subset of the initially identified features using techniques such as filter, wrapper, and embedded methods (Theng and Bhoyar, 2024; Krämer et al., 2023; Khaire and Dhanalakshmi, 2022; Palo et al., 2021; García-Magariño et al., 2020). The filter method employs statistical measures like correlation, while wrapper methods assess how features impact prediction accuracy through techniques such as backward, forward, stepwise selection, permutation feature importance (PFI), column dropout variants, SHAP, and Local Interpretable Model Explanations (LIME). Embedded methods explore how features affect prediction accuracy during training by leveraging inherent FI techniques. No matter what technique is used in feature selection, domain knowledge remains an intrinsic component guiding the process (Hoxha, 2024; Lawal Dano, 2023; Jha et al., 2020). Fig. presents a flowchart illustrating the DR process and its components.

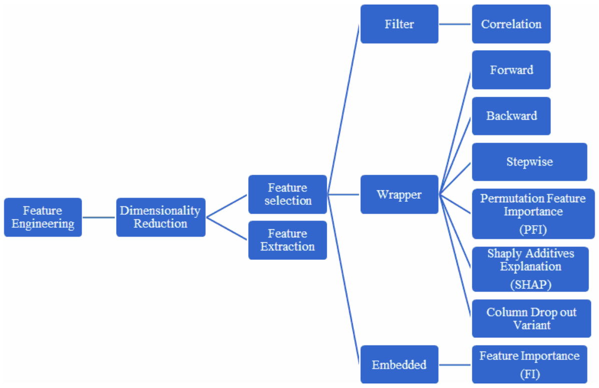


Fig. 1. The DR flowchart.

While FI techniques are typically employed to address the black-box aspect of machine learning models, with the goal of making them more explainable and interpretable, they can also be used for DR by selecting features with high FI, commonly known as top influential features. The FI can be model-specific or model-agnostic. Model-specific FI refers to the methods that can work within a certain model, such as those inherently coded within DT, RF, and XGBoost. In contrast, model-agnostic methods are those that can be plugged into any model, such as PFI, SHAP, and LIME. Some model-agnostic methods can show both the local and global effects of a particular feature, such as SHAP. Figure 2 illustrates the FI techniques.

Equally important to the FI techniques are the models in which these techniques are applied. These models are categorized into two types: parametric and nonparametric (Potrawa and Tetereva, 2022). Parametric models employ a predetermined functional relationship between the dependent and independent variables (Yang et al., 2024), often making presumptions such as normal distribution, while nonparametric models do not rely on any prior assumptions, but, instead, learn the relationship from the data. Numan and Yusoff (2024a) noted that parametric models include examples such as Linear Regression (LR) and Geographically Weighted Regression (GWR). In contrast, nonparametric models can be categorized into machine learning and deep learning approaches. The machine learning category comprises tree-based models like Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Adaptive Boosting (AdaBoost), Light Gradient Boosting Machine (LightGBM), and XGBoost, alongside other techniques such as K-nearest Neighbor (KNN) and Support Vector Machine (SVM). Deep learning methods, on the other hand, include Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). A schematic overview of these model classifications is presented in Figure 3.

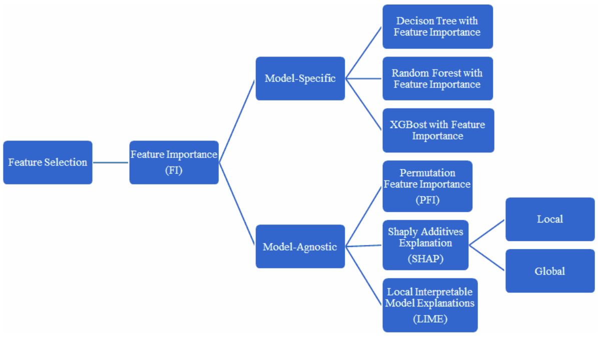

Feature Selection
Feature Importance (FI)
Model-Specific
Model-Agnostic
Decison Tree with Feature Importance
Random Forest with Feature Importance
XGBost with Feature Importance
Permutation Feature Importance (PFI)
Shaply Additives Explanation (SHAP)
Local Interpretable Model Explanations (LIME)
Local
Global

Fig. 2. The FI techniques flowchart.

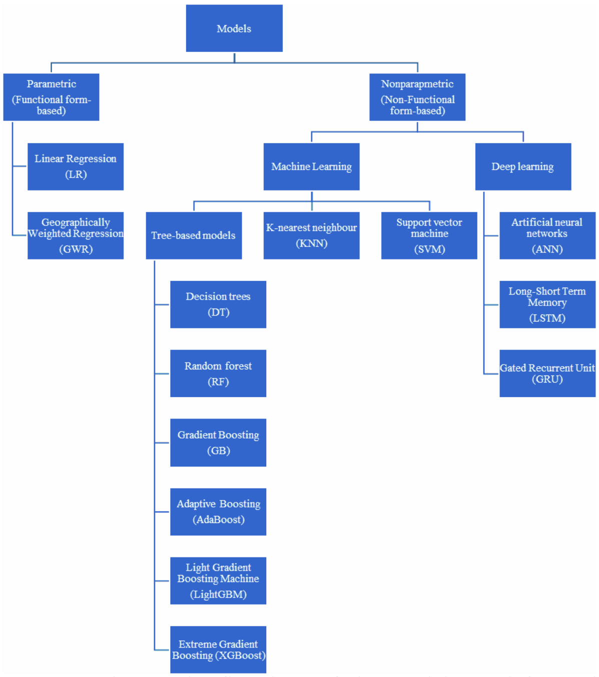

Models
Parametric (Functional form-based)
Nonparapmetric (Non-Functional form-based)
Linear Regression (LR)
Geographically Weighted Regression (GWR)
Machine Learning
Deep learning
Tree-based models
K-nearest neighbour (KNN)
Support vector machine (SVM)
Artificial neural networks (ANN)
Long-Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Decision trees (DT)
Random forest (RF)
Gradient Boosting (GB)
Adaptive Boosting (AdaBoost)
Light Gradient Boosting Machine (LightGBM)
Extreme Gradient Boosting (XGBoost)

Fig. 3. The flowchart of the models used for real estate appraisal.

To address the third research question, which pertains to evaluating model performance by comparing results when all features are incorporated versus using only the top influential features, this can be achieved by employing commonly used statistical metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2) (Numan and Yusoff, 2024a).

The structure of this paper is outlined as follows. Section 1 introduces the research, detailing its importance, issues related to model-based real estate appraisal methods, research aim, and research questions. Section 2 reviews relevant studies, focusing on empirical work that uses forms of DR through feature selection with FI techniques within tree-based models. This literature review identifies the research gap that distinguishes this study from others. Section 3 describes the methodology employed in the research. Section 4 provides an overview of the dataset. Section 5 presents the analysis, results, and discussion, while Section 6 concludes with a summary and final remarks.

2. LITERATURE REVIEW

In the field of real estate appraisal, extensive literature delves into the utilization of DR for model explainability by ranking the features that most significantly impact the value of the appraisal. However, a noticeable gap exists in research papers explicitly focused on investigating the effects of DR on model performance, specifically comparing the performance values when all the initially identified features are incorporated versus when only the most influential are included. It is worth mentioning that even when such research does exist, it often lacks implementing multiple techniques, models, and evaluation metrics. For example, while Hong et al. (2020) demonstrated improved performance in terms of the MAPE metric by selecting 16 features out of 26 using RF with FI on a dataset of 16,601 residential property transactions in Gangnam, South Korea, their study is limited to a single model, FI method, and evaluation metric.

In all cases, it is highly beneficial to explore research papers that implemented DR leveraging feature selection for interpretability purposes, in order to gain a close understanding of the settings and contexts in which these studies were conducted. This is explicitly highlighted by Theng and Bhoyar (2024), who emphasize that the choice of feature selection method heavily depends on dataset characteristics such as feature count, sample size, domain specificity, and statistical properties like central tendency, dispersion, skewness, outliers, correlation, and distribution.

Table 1 provides a summary of selected studies that implement DR through feature selection methods primarily aimed at enhancing model explainability. It includes details such as authors, publication dates, countries, dataset sizes, utilized models, feature selection methods, initially identified features, top influential features, and performance comparisons, if any, between incorporating all features versus only the top features.

Table 1. Summary of reviewed empirical studies employing DR based on feature selection, mainly for the purpose of interpretability.

No.	Study	Country	Data size	Model	Feature selection method	Number of initially identified features	Number of selected features	Performance comparison. All features versus only top features
1	Hong et al. (2020)	South Korea	16,601	RF	Inherent FI	26	16	One model, one FI method, and one evaluation metric
2	Iban (2022)	Turkey	1,002	RF, GBM, LightGBM, XGBoost, OLS	SHAP PFI	43	11	Not provided
3	Soltani et al. (2022)	Australia	428,000	GB	Inherent FI	38	10	Not provided
4	Rico-Juan and de La Paz (2021)	Spain	56,000	RF	SHAP	52	32	Not provided
5	Baur et al. (2023)	USA	33,610	GBM, RF, SVM, Elastic Net, and LR	SHAP	9	Not stated	Not provided
6	Li et al. (2021)	China	12,137	XGBoost LR	Inherent FI	35	5	Not provided
7	Aydinoglu and Sisman (2024)	Turkey	200,000	RF	Inherent FI	121	54	Not provided
8	Mete and Yomralioglu (2023)	UK	5,627,022	RF	PFI, SHAP	38	20	Not provided
9	Zaki et al. (2022)	USA	506	XGBoost	Inherent FI	13	Not stated	Not provided
10	Krämer et al. (2023)	German	81,166	XGBoost	PFI	Not stated	5	Not provided

Looking at the last column in Table 1, it is clear that previous studies that utilized some form of DR in the real estate appraisal industry did not explicitly assess its impact on model performance by comparing results from models that used all features against those that used only a subset of top influential features. When studies do make this comparison, they often fall short by not employing a variety of feature selection methods, models, and evaluation metrics. This research aims to fill this gap by conducting a comprehensive analysis that addresses these limitations.

Other than in the real estate domain, DR is also employed to improve performance across various fields. For example, in predicting the compressive strength of concrete, Wan et al. (2021) implemented DR by comparing the use of all features with features obtained through PCA and manual selection, processed using XGBoost, ANN, SVR, and LR models. The results indicate that DR improved performance, with XGBoost achieving higher R^2 and MSE values, ANN and LR performing better in MSE, and SVR showing improved R^2. This reinforces the notion of employing multiple statistical methods to assess performance. Similarly, Yang et al. (2023) applied DR to reduce the number of features affecting the mechanical properties of steel from 46 to 13, utilizing ANN and XGBoost models. Their findings indicated that the relative error between predicted and actual values was less than 5% when incorporating the full features into the models versus the reduced features. In the field of cybersecurity, Disha and Waheed (2022) used DR to reduce the number of features influencing intrusion detection systems from 42 to 20 in one set and from 41 to 10 in another. They evaluated performance using DT, AdaBoost, GBT, ANN, LSTM, and GRU models, comparing results from full-feature sets to reduced-feature sets. These studies significantly demonstrate the effect of DR in shaping performance, either by increasing it or maintaining it within an acceptable range, considering the advantages gained.

3. METHODOLOGY

The research methodology for implementing DR to measure its effect on improving real estate appraisal performance is guided by research questions that include identifying features influencing real estate appraisal, particularly for apartments within residential buildings in Al Bireh city; demonstrating how DR is applied through feature selection using multiple FI techniques across various models; and evaluating model performance, as elaborated in the following sections.

3.1. Features identification

Features recognized as influencing real estate appraisal significantly impact its performance (Aydinoglu and Sisman, 2024; Hoxha, 2024; Oust et al., 2023; Glumac and Des Rosiers, 2021).

Table 2. The set of 28 features initially identified as influencing real estate appraisal, together with their definitions and the appraisal value designated as the target variable, is presented within the context of Al Bireh city.

No.	Feature	Description
0	Appraisal value	The target variable representing the estimated value of an apartment, measured in Jordanian Dinar (JD). The exchange rate is 1 JD = 1.3 USD
Temporal		
1	Appraisal date	The date when the appraisal was conducted. The date type is an integer indicating the year of appraisal
2	Construction date	Represents the year the apartment was constructed. The values are expressed as integers
Physical		
3	Area	Represents the net surface area of the apartment in square meters (m²), excluding shared spaces like stairs and corridors. The values are expressed as real numbers
4	Number of apartments in building	The number of apartments within the building, expressed as an integer
5	Number of apartments in floor	The number of apartments on the same floor, expressed as an integer
6	Number of floors in building	The total number of floors in the building, expressed as an integer
7	Floor level	Indicates the specific floor level of the apartment. Values are integers, with negative values representing floors below ground
8	Floor to parcel ratio	The ratio of the apartment's floor area to the parcel area on which the building is constructed, expressed as a real number with three decimal places
9	Number of bedrooms	The number of bedrooms in the apartment, expressed as an integer
10	Number of bathrooms	The number of bathrooms with a shower area
11	Number of toilets	The number of toilets without a shower area
12	Number of balconies	The number of balconies, providing outdoor space and a view
13	Number of facades	The number of walls with windows in the apartment, ranging from 1 to 4
14	Central heating availability	Indicates whether central heating is available in the apartment (1 = yes, 0 = no)
15	Elevator availability	Indicates whether an elevator is available in the building (1 = yes, 0 = no)
16	Parking availability	Indicates whether a parking space is available (1 = yes, 0 = no)
17	Storage availability	Indicates whether a storage room is available (1 = yes, 0 = no)
18	Wall construction material	The material used in the apartment's walls (1 = masonry, 2 = concrete, 3 = bricks)
Surrounding and neighborhood		
19	Block	Indicates the neighborhood, with values representing the block number
20	Adjacent street type	Classification of the largest road adjacent to the building (1 = local street, 2 = main street)
21	Adjacent street width	The width of the largest road adjacent to the building in meter, measured as an integer
22	Number of adjacent streets	The number of streets adjacent to the building, expressed as an integer
Locational		
23	Proximity to hospitals	The distance to the nearest hospital, rated from 1 (shortest) to 5 (longest)
24	Proximity to schools	The distance to the nearest school, rated from 1 (shortest) to 5 (longest)
25	Proximity to city center	The distance to the city center, rated from 1 (shortest) to 5 (longest)
26	Proximity to main roads	The distance to the nearest main road, rated from 1 (shortest) to 5 (longest)
27	Proximity to area C	The distance to area C, rated from 1 (shortest) to 5 (longest). According to the Oslo Agreement of 1994, the lands in the West Bank are divided into three categories: area A, governed by the Palestinian Authority for both administrative and security affairs; area B, administratively controlled by the Palestinian Authority with security controlled by Israel; and area C, fully managed by Israel
28	Proximity to colonies	The distance to Israeli colonies, rated from 1 (shortest) to 5 (longest). The West Bank contains approximately 132 Israeli settlements, inhabited by over 700,000 Israeli settlers

The list of 28 features initially identified as influencing real estate appraisal, along with their definitions, references, and the appraisal value as the target variable in the context of Al Bireh city.

This study initially identifies 28 features affecting real estate appraisal from literature review as well as from local laws and regulations applicable to the study area, namely Al Bireh city.

These features, detailed in Table 2, along with their definitions and references, are categorized into four main groups: temporal, physical, surrounding and neighborhood, and locational.

The temporal features encompass information related to dates, such as the construction date and the appraisal date. Physical features represent the characteristics of the apartments, including area, the number of apartments in the building, the number of apartments on each floor, the number of floors in the building, floor level, floor to parcel ratio, the number of bedrooms, bathrooms, toilets, and balconies, the number of facades, and the availability of central heating, elevator, parking, and storage.

The surrounding and neighborhood features include attributes such as block, adjacent street type, adjacent street width, and the number of adjacent streets. Locational features cover proximity to hospitals, schools, city center, main roads, area C, colonies, and the appraisal value, which serves as the target variable. This set of potential features guides the data collection process to focus on aspects relevant to real estate appraisal.

3.2. Models and FI techniques

The second research question explores how to apply dimensionality reduction (DR) through the feature selection method using two feature importance (FI) techniques, namely inherent FI and SHAP, within three tree-based models—DT, RF, and XGBoost—as well as linear regression (LR) as a benchmark, with the aim of selecting the most influential features from those initially identified. This allows for a comparison between model performance when all initial features are included versus when only the most influential ones are. The two FI techniques with four models result in eight plots, each representing the ranking of features in descending order. To select the top influential features, each feature that falls within the top six in each plot is assigned a score of "1". The total score for each feature is then calculated by summing these individual scores across the eight plots. Features with the highest scores are considered the top influential features. The following subsections provide fundamental background on the models used and FI techniques.

3.2.1. LR

The LR serves as traditional reference model (Lenaers et al., 2024). It is constructed by determining the intercept and coefficients that minimize the prediction error, referred to as the residual, through the ordinary least squares approach. The LR prediction formula can be expressed in Equation 4, facilitating its application in predicting outcomes for any provided observation.

$$y = \beta_0 + \sum_{k=1}^{m} \beta_k x_k + \varepsilon \tag{1}$$

where:

y - predicted value (representing the real estate appraisal or predicted price);

β_0 - intercept that stands for independent features absent in the model (corresponding to the predicted value when all other features equal zero);

β_k - coefficient of the kth independent feature;

x_k - value of the observation for the kth independent feature;

ε - error;

m - total number of independent features.

3.2.2. DT

The DT training includes splitting the dataset into subsets such that the variance or MSE is minimized (Louati et al., 2022; Potrawa and Tetereva, 2022). The DT iteratively splits the subsets until no further improvements are achievable or until reaching specified constraints imposed by hyperparameters, such as maximum tree depth or minimum observations required in a node. The prediction can be written as shown in the equation below (Cohen et al., 2015):

$$\mathrm{f}(\mathrm{x}_i) \sum_{q}^{Q} C_q I_{R_q}(x_i) \tag{2}$$

where:

$f(x_i)$ - prediction of the ith observation;

x_i - ith observation;

C_q - constant;

I_{Rq} - indicator function of the qth subset of the dataset.

3.2.3. RF

The RF, originally introduced by Breiman (2001), consists of multiple DTs operating in parallel independently. Each tree is trained on a bootstrap sample, which comprises the same number of observations as the original dataset but not necessarily the same number of features. Although the bootstrap sample shares the same number of observations as the original dataset, it differs in that it allows for the repetition of observations, a process known as sampling with replacement. Each individual tree within the RF behaves similarly to a standard single DT as discussed previously. The final prediction for an observation is simply the average of the predicted values from each tree. This prediction can be mathematically expressed using the equation adopted by Potrawa and Tetereva (2022):

$$\mathrm{f}(\mathrm{x}) = \frac{\sum_{p=1}^{P} f_p(X_p)}{P} \tag{3}$$

where:

f(x) - prediction of the x observation;

f_p - prediction of the x^{th} observation in the p^{th} tree;
X_p - features of the p^{th} tree;
P - total number of trees.

3.2.4. XGBoost

The XGBoost, developed by Chen and Guestrin (2016), consists of multiple DTs operating sequentially dependently. In the first tree, the XGBoost is trained on the target observations while in the subsequent trees, the XGBoost is trained on the residuals (Lorenz et al., 2023). Each individual tree within the XGBoost behaves similarly to a standard single DT as discussed previously in terms of splitting and minimizing the loss functions. The final prediction for an observation is simply the predicted values from each tree multiplied by the learning rate (η). This prediction can be mathematically expressed using the equation adopted by Sibindi et al. (2023):

$$f(x) = \eta \sum_{p=1}^{P} f_p(X_p) \tag{4}$$

where:
f(x) - prediction for the x^{th} observation;
η - learning rate;
f_p - prediction of the x^{th} observation made by the p^{th} tree;
X_p - features used by the p^{th} tree;
P - total number of trees (Zippo et al., 2025).

3.2.5. The inherent FI

In the case of a single decision tree, during training, all features of the dataset are explored to select the feature and threshold that minimizes the variance, thus creating the first node and data split. The variance is mathematically expressed as follows:

$$\sigma^2 = \frac{\sum(y_i - \mu)^2}{n} \tag{5}$$

where:
σ^2 - variance;
y_i - actual value of the observation;
μ - average;
n - number of observations.

The reduction in variance at each node is calculated by the weighted sum of variances of the subsets. This calculation is expressed mathematically as Equation (1), which originates from the information gain equation adopted by Soltani et al. (2022), representing the weighted sum of variance:

$$FI_j^k = \frac{n_l}{n}\sigma_l^2 + \frac{n_r}{n}\sigma_r^2 \tag{6}$$

where:
FI^k_d - importance (score) of the k^{th} feature split at j^{th} node;
n_l - number of observations of the subset split at left;
σ^2_l - variance of left subset;
n_r - number of observations of the subset split at right;
σ^2_r - variance of the subset in the right split;
n - total number of observations in the left and right subsets.

The overall FI in a single tree is calculated by aggregating its weighted sum of variances at all nodes in the tree. The feature with the highest aggregated weighted sum of variances across all nodes of the tree is considered to be more important than others. In the case of RF, the FI is calculated by aggregating the weighted sum of variances across all nodes in all trees. The same approach is followed when it comes to XGBoost. In the case of LR, the FI plot is based on the coefficients, where the magnitude indicates the effect of a feature on the prediction.

3.2.6. SHAP

The SHAP technique, developed by Lundberg and Lee (2017), is model-agnostic used to assess the individual contributions of features to a model's prediction. The SHAP value for a specific feature is determined by summing the weighted differences in predictions for all possible subsets of features, both including and excluding that particular feature. In straightforward terms, the SHAP value pertaining to a specific feature indicates the average prediction adjustment across all possible combinations involving that feature and other features, where the unit of the SHAP value corresponds to the unit of the target variable (Kraus et al., 2020).

Mathematically, the SHAP formula is represented as follows:

$$\varphi_{k=} \sum_{q=1}^{Q} \frac{|s_q|!\,(m! - |s_q| - 1)!}{m!}\left[f(X_{s_q \cup k}) - f(X_{s_q \setminus k})\right] \tag{7}$$

where:
φk - SHAP value of the kth feature;
sq - qth subset of feature;
|sq| - number of features in the qth subset;
"!" denotes the mathematical factorial;
Xsq∪k - data containing the qth subset of features including the k feature;
Xsq\k - data containing the qth subset of features excluding the k feature;
f - model;
q - subset of features;
Q - number of all possible feature subsets for the kth feature;
m - total number of features.

3.3. Performance evaluation

In this research, the most commonly used model performance evaluation metrics in the field of real estate appraisal are employed, specifically MSE, RMSE, MAE, MAPE, and R² (Numan and Yusoff, 2024a). While MSE, RMSE, MAE, and MAPE metrics measure errors in different ways, R² evaluates the goodness-of-fit. Their definitions, along with their mathematical equations, are presented in Table 1.

Table 3. The set of statistical metrics selected to assess the performance of the hybrid model.

Metric	Definition	Equation
Mean Squared Error (MSE)	This metric involves averaging the square of the errors, magnifying larger errors while minimizing smaller ones. This means that its value is strongly affected by outliers. Its unit is meaningless. The closer the value to zero, the better.	$\frac{\sum_{i=1}^{n} e_i{}^2}{n}$ $e = Y_i - \hat{Y}_i$
Root Mean Squared Error (RMSE)	This measure involves taking the square root of the MSE to bring it back to the scale of the dependent feature. It has the same unit as the target feature. The closer the value to zero, the better.	$\sqrt{\frac{\sum_{i=1}^{k} e_i{}^2}{n}}$
Mean Absolute Error (MAE)	This metric calculates the average of absolute errors, assigning equal weight to all errors. As a result, it is not influenced by outliers. A value closer to zero indicates better performance.	$\frac{\sum_{i=1}^{n} \lvert e \rvert}{n}$
Mean Absolute Percentage Error (MAPE)	This measure involves averaging the absolute error relative to the actual value, leading to different treatment for errors of similar magnitude. The closer the value to zero, the better.	$\frac{\sum_{i=1}^{k} \left(\frac{\lvert e_i \rvert}{Y_i}\right)}{n}$
Coefficient of Determination (R²)	This metric measures the percentage of the squared sum of errors relative to the squared sum of errors in the mean, subtracted from 1. The closer the value to 1, the better. It indicates the percentage of the variation in the target variable that can be explained by the features in the model.	$1 - \frac{\sum_{1}^{n} e_i{}^2}{\sum_{1}^{n} (Y_i - \bar{Y}_i)^2}$

**** ei is the error in the prediction of the ith observation which is the difference between the actual value and predicted value, Yi is the actual value of the ith observation of the dependent feature, Ŷi is the predicted value of the ith observation of the dependent feature, Ȳ is the average of the actual values of the dependent feature for all observations, and n is the number of observations (source: Steurer et al., 2021).*

4. DATA

Data collection is guided by the 28 features initially identified as influencing the appraisal of residential apartments within buildings in Al Bireh city, as presented in Table 2. Within this study area, the data for these features are not consolidated in a single agency but is available across three main sources: Al Bireh Municipality (BM), Palestine Land Authority (PLA), and the Ministry of Local Government (MOLG).

In the first data collection phase, information for 21 features are gathered from BM (construction date, area, number of apartments in the building, number of apartments on each floor, number of floors in the building, floor level, floor-to-parcel ratio, number of bedrooms, number of bathrooms, number of toilets, number of balconies, number of facades, central heating availability, elevator availability, parking availability, storage availability, wall construction material, block, adjacent street type, adjacent street width, and number of adjacent streets). As these data are not structured in a tabular format, they are extracted manually from the municipality's engineering plans database. A random sampling strategy is used to select records from engineering plans for residential buildings in the study area that are designated as appraised. Other details, such as parcel and quarter numbers, are also entered to be used as identifiers to link the data with other sources, while x and y coordinates are recorded for mapping, resulting in a dataset of 5.586 entries stored in an Excel file.

The second phase focuses on collecting two features-appraisal value and appraisal date-for each residential apartment. These two pieces of information are sourced from PLA and found to be documented in separate Microsoft Word files, which are integrated into the Excel file based on common parcel, quarter, and block numbers. Appraisal values are available for 2.354 observations, while the remaining records were excluded due to the absence of appraisal values in PLA. By the end of this phase, data collection for 22 features is complete, with the appraisal value as the target variable.

The third phase aims to gather the remaining six locational features related to proximity to amenities, including hospitals, schools, the city center, main roads, Area C, and colonies. These data are downloaded from the Geomolg Geoportal (geomolg.ps), managed by MOLG. Geographic Information Systems (GIS) techniques are used to derive the values for these six features, so they can be incorporated into the Excel data table.

At the conclusion of this phase, the dataset includes all 28 features, with the appraisal value as the target variable, comprising 2,354 observations. As part of the pre-processing step, invalid entries and missing data are corrected by revisiting the sources, either BM or PLA, to obtain the necessary corrections, while duplicate entries are removed. Based on the central limit theorem, checking for a normal distribution of the appraisal value as a target variable is unnecessary, given that the dataset contains more than 30 observations (Koohpayma and Argany, 2021). The descriptive statistics for the features are shown in Table 4.

Table 4. The Descriptive statistics of the 28 features identified as affecting real estate appraisal within the context of Al Bireh city.

No.	Feature	Minimum	Maximum	Mean	Standard Deviation
1	Appraisal value	30.961	79.992	63.986	10.401
2	Appraisal date	2008	2023	2016	3
3	Construction date	1995	2022	2011	5.33
4	Area	91	207	146	20
5	Number of apartments in building	2	38	17.05	6.22
6	Number of apartments in floor	1	5	2.62	0.80
7	Number of floors in building	2	11	6.69	1.47
8	Floor level	-6	8	1.68	2.13
9	Floor to parcel ratio	0.069	0.968	0.45	0.09
10	Number of bedrooms	1	5	2.96	0.23
11	Number of bathrooms	1	3	1.76	0.44
12	Number of toilets	0	2	0.74	0.45
13	Number of balconies	0	4	1.39	0.69
14	Number of facades	1	4	2.58	0.54
15	Central heating availability	0	1	0.02	0.14
16	Elevator availability	0	1	0.95	0.22
17	Parking availability	0	1	0.91	0.29
18	Storage availability	0	1	0.13	0.33
19	Wall construction material	1	2	1.00	0.07
20	Block	7	28	8	18
	Adjacent street type	1	4	2.27	0.64
21	Adjacent street width	3	30	10.30	3.76
22	Number of adjacent streets	1	3	1.45	0.60
23	Proximity to hospitals	1	5	2.37	1.02
24	Proximity to schools	1	4	1.70	0.70
25	Proximity to city center	1	5	3.53	0.90
26	Proximity to main roads	1	4	1.51	0.80
27	Proximity to area C	1	4	2.12	1.03
28	Proximity to colonies	1	5	3.25	0.91

5. ANALYSIS, RESULTS, AND DISCUSSION

The dataset, consisting of 28 features and 2.345 observations, including the appraisal value as the target variable and meeting the pre-processing requirements, is fed into the three tree-based models and the LR using open-source Python libraries such as pandas, sklearn, numpy, matplotlib, and scikit-learn, within JupyterLab in the Anaconda App. Although hyperparameter values are typically selected from a set of candidates using strategies that optimize model performance, such as grid search, random search, and Bayesian, in this research, the hyperparameter values for DT, RF, and XGBoost are assumed as depicted in Table 5. These values are guided by ranges observed in similar studies within the field of real estate appraisal (Mathotaarachchi et al., 2024; Neves et al., 2024; Sharma et al., 2024; Choy and Ho, 2023; Hu and Tang, 2023; Sibindi et al., 2023; Zhang et al., 2023; Kim et al., 2022; Louati et al., 2022; Kim et al., 2021).

Table 5. The selected hyperparameters for the DT, RF, and XGBoost.

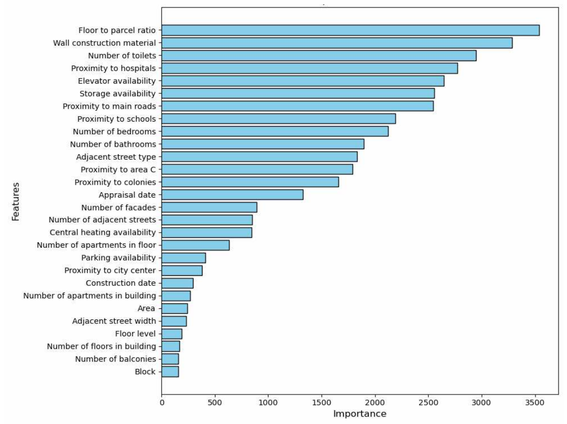
Hyperparameter	Selected value		
	DT	RF	XGBoost
min_samples_split	2	5	---
min_samples_leaf	1	5	---
max_depth	5	10	5
n_estimators	---	100	100
subsample	---	---	1
colsample _bytree	---	---	0.9
Learning_rate (eta)	---	---	0.2
lambda	---	---	0.2
alpha	---	---	10

The result of three tree-based machine learning models: DT, RF, and XGBoost, with the hyperparameters listed in Table 5, along with the 5-fold cross validation, and considering five statistical performance evaluation metrics presented in Table 3 are summarized in Table 6. The results clearly indicate that the XGBoost has the best performance in all the metrics.

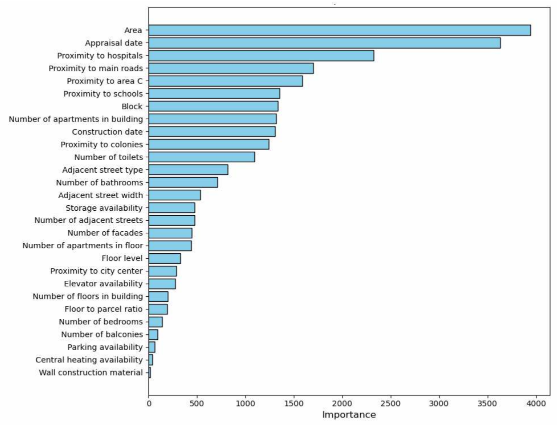
Table 6. The performance evaluation metrics for LR, DT, RF, and XGBoost, when 28 features are incorporated along with the appraisal value as the target variable.

Model	MSE	RMSE	MAE	MAPE	R²
LR	54,304,379	7,354	5,202	8.969	0.497
DT	42,764,662	6,539	4,690	7.944	0.604
RF	21,848,492	4,658	2,533	4.411	0.798
XGBoost	19,941,336	4,465	2,061	3.530	0.815

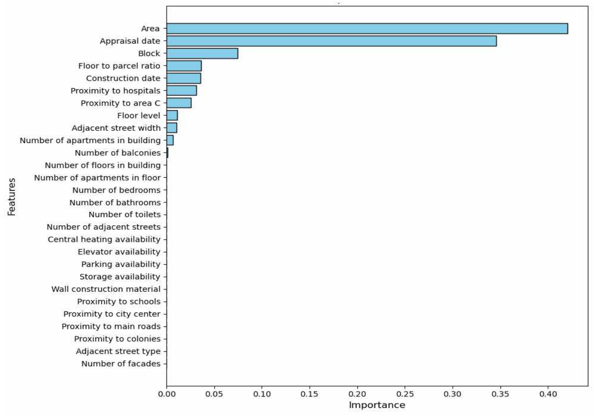
LR (Coefficients)



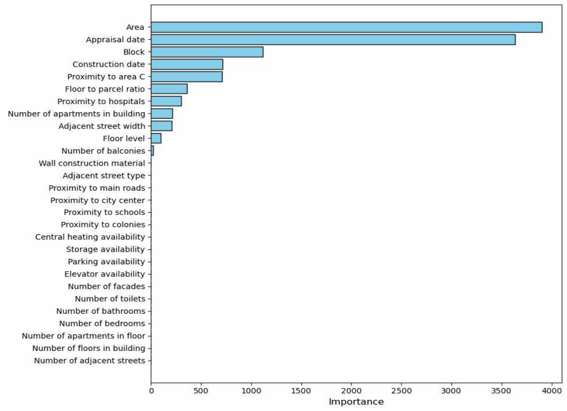
LR (SHAP)



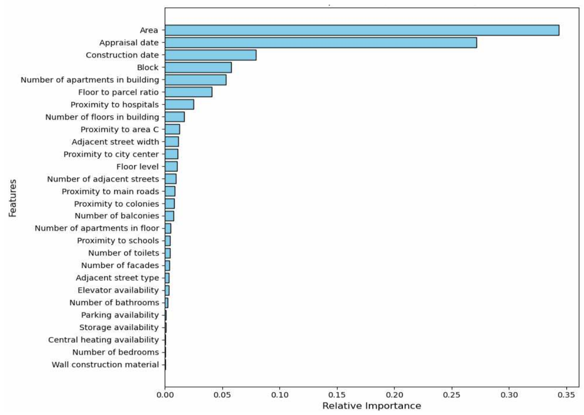
DT (inherent FI)



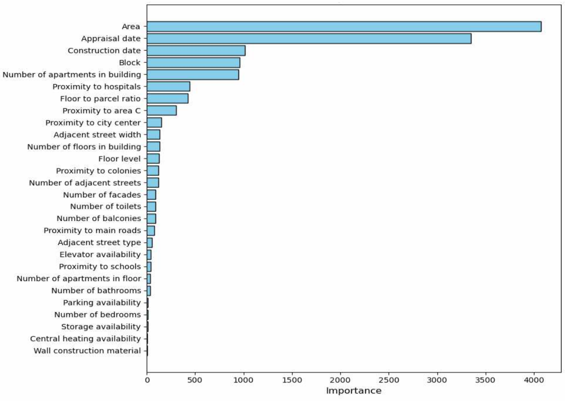
DT (SHAP)



RF (inherent FI)



RF (SHAP)



XGBoost (inherent FI)

XGBoost (SHAP)

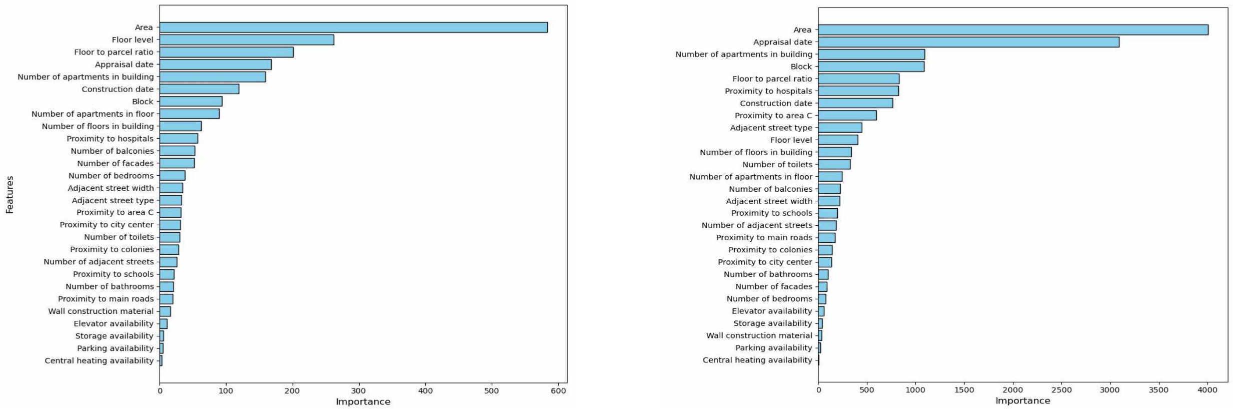


Fig. 4. The results of inherent FI and SHAP within the DT, RF, and XGBoost, based on 28 features with the appraisal as the target variable. The coefficients are plotted for LR as an equivalent method for inherent FI.

While Soltani et al. (2022) implemented top influential features based on the model with the highest performance, this research applies two FI techniques across four models. Specifically, inherent FI and SHAP techniques are employed to rank features in descending order of importance, allowing for the selection of the top six features out of the initially identified 28. In the case of LR, inherent FI is expressed by plotting the feature coefficients, indicating the importance of each feature in determining the target variable. The plots are shown in Figure 4. It presents eight plots of ranked features derived from four models, each using two different FI techniques. To obtain a unified set of the top six influential features, each feature that falls within the top six is assigned a score of "1", as shown in Table 7.

Table 7. The total score of the top six most influential features using inherent FI and SHAP within the DT, RF, XGBoost, and LR.

No.	Feature	Inherent FI				SHAP				Total score
		LR	DT	RF	XG Boost	LR	DT	RF	XG Boost	
1	Appraisal date		1	1	1	1	1	1	1	7
2	Area		1	1	1	1	1	1	1	7
3	Construction date		1	1			1	1		4
4	Block		1	1			1	1	1	5
5	Number of apartments in building			1	1			1	1	4
6	Number of apartments in floor									
7	Number of floors in building									
8	Floor level				1					1
9	Floor to parcel ratio	1	1		1				1	4
10	Number of bedrooms									
11	Number of bathrooms									
12	Number of toilets	1								1
13	Number of balconies									
14	Number of facades									
15	Central heating availability									
16	Elevator availability	1								1
17	Parking availability									
18	Storage availability									
19	Adjacent street type									
20	Adjacent street width									
21	Number of adjacent streets									
22	Wall construction material	1								1
23	Proximity to hospitals	1				1				2
24	Proximity to schools									
25	Proximity to city center									
26	Proximity to main roads					1				1
27	Proximity to area C					1	1			2
28	Proximity to colonies									

The total score for each feature is then calculated by summing these individual scores across the eight plots, as represented in the last column of Table 7. According to this column, the six features with the highest scores are appraisal date, area, construction date, block, number of apartments in the building, and floor to parcel ratio. The selection of the top six features is based on their high total scores, which range from 7 to 4, before dropping to 1 or less for the remaining features. The identification of the appraisal date, construction date, and area among the top six features aligns with the results obtained by Krämer et al. (2023) through their research utilizing the PFI within XGBoost. Similarly, this finding is supported by Aydinoglu and Sisman (2024), who identified construction date and area among the top three features using RF with the inherent FI. The next major step in the analysis is evaluating the performance by inputting these selected top six features into the four models with identical hyperparameters and a 5-fold cross-validation technique. The outcomes of the statistical performance evaluation metrics are consolidated in Table 8.

Table 8. The performance evaluation metrics for LR, DT, RF, and XGBoost, considering the top six features with the appraisal value as the target variable.

Model	MSE	RMSE	MAE	MAPE	R^2
LR	63,490,393	7,959	5,925	10.2	0.412
DT	43,883,329	6,624	4,798	8.13	0.594
RF	22,422,585	4,718	2,629	4.578	0.793
XGBoost	20,950,174	4,577	2,124	3.658	0.806

To facilitate a comprehensive comparison of the performance of the four models using both all initial features (Table 2) and only the top six influential features (Table 7), the relative performance change is calculated. This is done by subtracting the performance metric value when all features are included from the performance metric value when only the top six features are included, then dividing the result by the performance metric when all features are included, as presented in Figure 5.

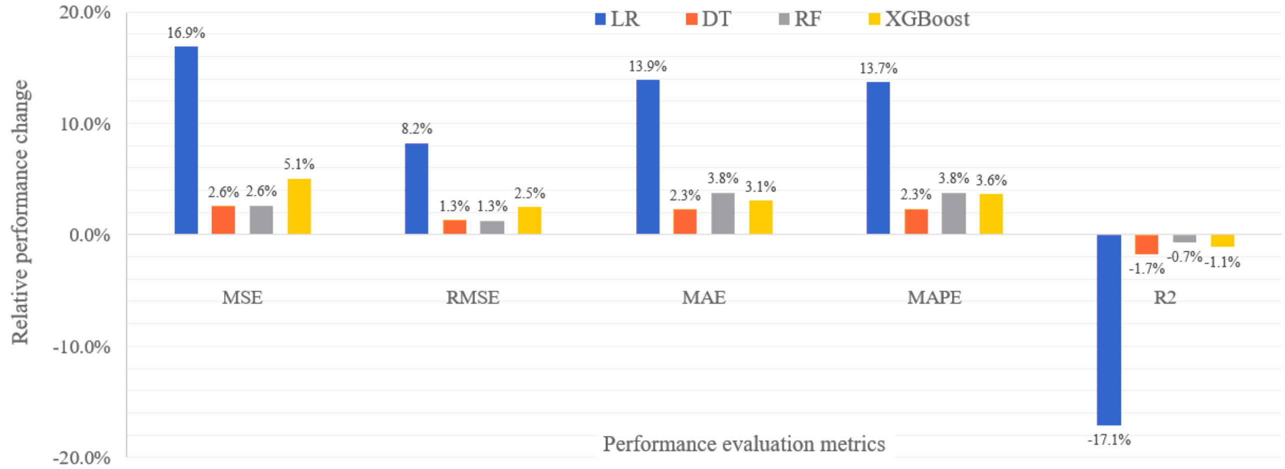


Fig. 5. The relative performance change when using all features versus using the top six features.

The results clearly demonstrate that none of the four models achieves any improvement across any of the five metrics. This indicates that the performance with all features included in the models is still better than when only the most influential features are included. However, with the exception of the LR model, the relative performance change remains below approximately 5% and is as low as 0.7% in terms of R^2 for the RF model. These results align with the findings of Yang et al. (2023), where the relative error between the actual and predicted values when using the full features and the reduced features was less than 5%. This raises the question of whether the trade-off between model performance and reduced computational time and hardware resources is worthwhile. A similar trade-off is also considered between performance and interpretability (Kucklick and Müller, 2023) or between performance and stability (Khaire and Dhanalakshmi, 2022), to avoid focusing solely on one aspect while neglecting others.

6. CONCLUSIONS

This study highlights the significant role of DR in eliminating redundancy, irrelevance, noise, and complexity, thereby reducing computational time, hardware resources, storage requirements, and data collection efforts. With these advantages in mind, the research aimed to explicitly investigate the potential of DR to improve real estate appraisal performance by reducing the number of features through a feature selection method, employing two FI techniques across four models, and relying on five statistical metrics for model performance evaluation.

Concerning the first research question related to the identification of the initial features influencing real estate appraisal, particularly residential apartment buildings in Al Bireh city, these features are identified based on the literature, local laws, and regulations, and are presented in Table 2. The dataset includes 28 features along with the appraisal value as the target variable, with a preprocessed dataset consisting of 2,354 observations for residential properties. Descriptive statistics for the features are shown in Table 4.

The second research question, focusing on conducting DR through selecting the most influential features, employs multiple FI techniques (inherent FI and SHAP techniques) and multiple models (DT, RF, XGBoost, and LR). The two FI methods, both inherent FI and SHAP, are conducted within the four models, yielding eight plots of ranked features as illustrated Figure 4. The initial 28 features are summarized in a table, where each feature that lies within the top six according to the FI plots and its corresponding model is assigned a score of “1”. The total score for each feature is then calculated. The results indicate that the features with the highest total scores are six: appraisal date, area, construction date, block, number of apartments in the building, and floor-to-parcel ratio.

To address the third question, separately, all the 28 initial features and the top six influential ones are fed into the 4 models along with the assumed hyperparameters and 5-fold cross-validation training technique. The five statistical performance evaluation metrics - MSE, RMSE, MAE, MAPE, and R² - are determined. The findings indicate that when only the top six features were used, none of the models achieved performance improvement compared to including all features. However, the relative performance metric change warrants interpretation, as, except for LR, it was less than approximately 5% for all models and as low as 0.7% in terms of R² for RF. This probably suggests a need to balance performance with computational time, hardware resources, and data collection. This trade-off echoes the traditional debate between performance and interpretability, or performance and stability, reinforcing the idea that performance does not necessarily come at the expense of other equally important factors.

Several limitations need to be mentioned. The features analyzed in this study only encompass the physical and locational attributes of the apartments, without considering microeconomic factors due to the lack of such data. For the same reason, the dataset size is relatively small for machine learning models. Larger datasets increase the likelihood of models being trained on diverse cases, enabling them to learn more patterns effectively. In addition to the importance of dataset size, representativeness also plays a significant role. Despite these limitations, this research makes a valuable contribution by presenting an empirical investigation into DR through feature selection, employing two FI techniques across four models.

The implications are threefold: Firstly, it provides insight into the magnitude of each feature’s impact on appraisal performance, enhancing explainability. Secondly, it offers a list of top features influencing apartment appraisals, which can benefit stakeholders such as apartment sellers and buyers, banks, property taxation authorities, insurance agencies, and those making informed development and investment decisions. Thirdly, it reduces the need for hardware resources, prolonged processing time, storage, and data collection during model training by eliminating redundancy, irrelevance, noise, and complexity through DR.

To better understand the potential impact of DR on model performance improvement, additional future empirical studies are needed. These studies should not be limited to tree-based machine learning models but should encompass a broader range of models with larger datasets in terms of both features and observations. On the other hand, even when using the same models – LR, DT, RF, and XGBoost – an alternative approach could assess performance by allowing each model to use its own top features, without relying on the total score method. This approach might have the potential to improve the performance of these models by focusing on their specific top features; however, it would reduce the likelihood of identifying a unified set of top influential features.

REFERENCES

Alzain E., Alshebami A. S., Aldhyani T. H., Alsubari S. N. (2022), Application of artificial intelligence for predicting real estate prices: the case of Saudi Arabia. Electronics, 11, 3448. DOI: https://doi.org/10.3390/electronics11213448

Aydinoglu A. C., Sisman S. (2024), Comparing modelling performance and evaluating differences of feature importance on defined geographical appraisal zones for mass real estate appraisal. Spatial Economic Analysis, 19, 225-249. DOI: https://doi.org/10.1080/17421772.2023.2242897

Baur K., Rosenfelder M., Lutz B. (2023), Automated real estate valuation with machine learning models using property descriptions. Expert Systems with Applications, 213, 119147. DOI: https://doi.org/10.1016/j.eswa.2022.119147

Breiman L. (2001), Random forests. Machine learning, 45, 5-32. DOI: https://doi.org/10.1023/A:1010933404324

Chanasit K., Chuangsuwanich E., Suchato A., Punyabukkana P. (2021), A real estate valuation model using boosted feature selection. IEEE Access, 9, 86938-86953. DOI: 10.1109/ACCESS.2021.3089198

Chen T., Guestrin C. (2016), Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794. DOI: https://doi.org/10.1145/2939672.293978

Chen Y., Yang Q., Geng L., Yin W. (2024), Analysis of Factors Influencing Housing Prices in Mountain Cities Based on Multiscale Geographically Weighted Regression—Demonstrated in the Central Urban Area of Chongqing. Land, 13, 602. DOI: https://doi.org/10.3390/land13050602

Chhikara P., Jain N., Tekchandani R., Kumar N. (2022), Data dimensionality reduction techniques for Industry 4.0: Research results, challenges, and future research directions. Software: Practice and Experience, 52, 658-688. DOI: https://doi.org/10.1002/spe.2876

Choy L. H., Ho W. K. (2023), The use of machine learning in real estate research. Land, 12, 740. DOI: https://doi.org/10.3390/land12040740

Çılgın C., Gökçen H. (2023), Machine learning methods for prediction real estate sales prices in Turkey. Revista de la construcción, 22, 163-177. DOI: http://dx.doi.org/10.7764/rdlc.22.1.163

Cohen S. N., Elliott R. J., Cohen S. N., Elliott R. J. (2015), Measure and Integral. Stochastic Calculus and Applications, 3-47. DOI: https://doi.org/10.1007/978-1-4939-2867-5

Das S. S. S., Ali M. E., Li Y.-F., Kang Y.-B., Sellis T. (2021), Boosting house price predictions using geo-spatial network embedding. Data Mining and Knowledge Discovery, 35, 2221-2250. DOI: https://doi.org/10.1007/s10618-021-00789-x

Disha R. A., Waheed S. (2022), Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. Cybersecurity, 5, 1. DOI: https://doi.org/10.1186/s42400-021-00103-8

Droj G., Kwartnik-Pruc A., Droj L. (2024), A Comprehensive Overview Regarding the Impact of GIS on Property Valuation. ISPRS International Journal of Geo-Information, 13, 175. DOI: https://doi.org/10.3390/ijgi13060175

Elnaeem Balila A., Shabri A. B. (2024), Comparative analysis of machine learning algorithms for predicting Dubai property prices. Frontiers in Applied Mathematics and Statistics, 10, 1327376. DOI: https://doi.org/10.3389/fams.2024.1327376

García-Magariño I., Medrano C., Delgado J. (2020), Estimation of missing prices in real-estate market agent-based simulations with machine learning and dimensionality reduction methods. Neural Computing and Applications, 32, 2665-2682. DOI: https://doi.org/10.1007/s00521-018-3938-7

Geerts M., De Weerdt J. (2023), A survey of methods and input data types for house price prediction. ISPRS International Journal of Geo-Information, 12, 200. DOI: https://doi.org/10.3390/ijgi13060175

Glumac B., Des Rosiers F. (2021), Practice briefing–Automated valuation models (AVMs): their role, their advantages and their limitations. Journal of Property Investment & Finance, 39, 481-491. DOI: https://doi.org/10.1108/JPIF-07-2020-0086

Hoang D., Wiegratz K. (2023), Machine learning methods in finance: Recent applications and prospects. European Financial Management, 29, 1657-1701. DOI: https://doi.org/10.1111/eufm.12408

Hong J., Choi H., Kim W. S. (2020), A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. International Journal of Strategic Property Management, 24, 140-152. DOI: https://doi.org/10.3846/ijspm.2020.11544

Hoxha V. (2024), Comparative analysis of machine learning models in predicting housing prices: a case study of Prishtina's real estate market. International Journal of Housing Markets and Analysis. DOI: https://doi.org/10.1108/IJHMA-09-2023-0120

Hu G., Tang Y. (2023), GERPM: A Geographically Weighted Stacking Ensemble Learning-Based Urban Residential Rents Prediction Model. Mathematics, 11, 3160. DOI: https://doi.org/10.3390/math11143160

Hurley A. K., Sweeney J. (2024), Irish property price estimation using a flexible geo-spatial smoothing approach: What is the Impact of an address? The Journal of Real Estate Finance and Economics, 68, 355-393. DOI: https://doi.org/10.1007/s11146-022-09888-y

Iban M. C. (2022), An explainable model for the mass appraisal of residences: The application of tree-based Machine Learning algorithms and interpretation of value determinants. Habitat International, 128, 102660. DOI: https://doi.org/10.1016/j.habitatint.2022.102660

Jha S. B., Babiceanu R. F., Pandey V., Jha R. K. (2020), Housing market prediction problem using different machine learning algorithms: A case study. arXiv preprint arXiv:2006.10092. DOI: https://doi.org/10.48550/arXiv.2006.10092

Jia W., Sun M., Lian J., Hou S. (2022), Feature dimensionality reduction: a review. Complex & Intelligent Systems, 8, 2663-2693. DOI: https://doi.org/10.1007/s40747-021-00637-x

Jin S., Zheng H., Marantz N., Roy A. (2024), Understanding the effects of socioeconomic factors on housing price appreciation using explainable AI. Applied Geography, 169, 103339. DOI: https://doi.org/10.1016/j.apgeog.2024.103339

Khaire U. M., Dhanalakshmi R. (2022), Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences, 34, 1060-1073. DOI: https://doi.org/10.1016/j.jksuci.2019.06.012. https://doi.org/10.3390/su14159056

Kim J., Lee Y., Lee M.-H., Hong S.-Y. (2022), A comparative study of machine learning and spatial interpolation methods for predicting house prices. Sustainability, 14, 9056. DOI: https://doi.org/10.3390/su14159056

Kim J., Won J., Kim H., Heo J. (2021), Machine-learning-based prediction of land prices in Seoul, South Korea. Sustainability, 13, 13088. DOI: https://doi.org/10.3390/su132313088

Koohpayma J., Argany M. (2021), Estimating the price of apartments in Tehran using extracted compound variables. International Journal of Housing Markets and Analysis, 14, 569-595. DOI: https://doi.org/10.1108/IJHMA-05-2020-0050

Krämer B., Nagl C., Stang M., Schäfers W. (2023), Explainable AI in a real estate context–exploring the determinants of residential real estate values. Journal of Housing Research, 32, 204-245. DOI: https://doi.org/10.1080/10527001.2023.2170769

Kraus M., Feuerriegel S., Oztekin A. (2020), Deep learning in business analytics and operations research: Models, applications and managerial implications. European Journal of Operational Research, 281, 628-641. DOI: https://doi.org/10.1016/j.ejor.2019.09.018

Kucklick J.-P., Müller O. (2023), Tackling the accuracy-interpretability trade-off: Interpretable deep learning models for satellite image-based real estate appraisal. ACM Transactions on Management Information Systems, 14, 1-24. DOI: https://doi.org/10.1145/356743

Lahmiri S., Bekiros S., Avdoulas C. (2023), A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization. Decision Analytics Journal, 6, 100166. DOI: https://doi.org/10.1016/j.dajour.2023.100166

Lawal Dano U. (2023), Analyzing the spatial determinants of housing prices in Dammam, Saudi Arabia: an AHP approach. International Journal of Housing Markets and Analysis. DOI: https://doi.org/10.1108/IJHMA-07-2023-0101

Lee C. (2021), Predicting land prices and measuring uncertainty by combining supervised and unsupervised learning. International Journal of Strategic Property Management, 25, 169-178. DOI: https://doi.org/10.3846/ijspm.2021.14293

Lenaers I., Boudt K., De Moor L. (2024), Predictability of Belgian residential real estate rents using tree-based ML models and IML techniques. International Journal of Housing Markets and Analysis, 17, 96-113. DOI: https://doi.org/10.1108/IJHMA-11-2022-0172

Li S., Jiang Y., Ke S., Nie K., Wu C. (2021), Understanding the effects of influential factors on housing prices by combining extreme gradient boosting and a hedonic price model (XGBoost-HPM). Land, 10, 533. DOI: https://doi.org/10.3390/land10050533

Lorenz F., Willwersch J., Cajias M., Fuerst F. (2023), Interpretable machine learning for real estate market analysis. Real estate economics, 51, 1178-1208. DOI: https://doi.org/10.1111/1540-6229.12397

Louati A., Lahyani R., Aldaej A., Aldumaykhi A., Otai S. (2022), Price forecasting for real estate using machine learning: A case study on Riyadh city. Concurrency and Computation: Practice and Experience, 34, e6748. DOI: https://doi.org/10.1002/cpe.6748

Lundberg S. M., Lee S.-I. (2017), A unified approach to interpreting model predictions. Advances in neural information processing systems, 30. DOI: https://doi.org/10.48550/arXiv.1705.07874

Mallick S., Mittal M. (2025), AI-Based Model Order Reduction Techniques: A Survey. Archives of Computational Methods in Engineering, 1-26. DOI: https://doi.org/10.1007/s11831-024-10207-2

Mankad M. D. (2022), Comparing OLS based hedonic model and ANN in house price estimation using relative location. Spatial Information Research, 30, 107-116. DOI: https://doi.org/10.1007/s41324-021-00416-3

Mathotaarachchi K. V., Hasan R., Mahmood S. (2024), Advanced Machine Learning Techniques for Predictive Modeling of Property Prices. Information, 15, 295. DOI: https://doi.org/10.3390/info15060295

Mete M. O., Yomralioglu T. (2023), A Hybrid Approach for Mass Valuation of Residential Properties through Geographic Information Systems and Machine Learning Integration. Geographical Analysis, 55, 535-559. DOI: https://doi.org/10.1111/gean.12350

Neves F. T., Aparicio M., De Castro Neto M. (2024), The impacts of open data and eXplainable AI on real estate price predictions in smart cities. Applied Sciences, 14, 2209. DOI: https://doi.org/10.3390/app14052209

Numan J. A., Yusoff I. M. (2024a), Identifying the current status of real estate appraisal methods. Real Estate Management and Valuation, 32, 12-27. DOI: https://doi.org/10.2478/remav-2024-0032.

Numan J. A., Yusoff I. M. (2024b), Identifying the key variables affecting condominium real estate appraisal in the context of Al Bireh City, Palestine: a literature and questionnaire-based approach. Journal of Facilities Management. DOI: https://doi.org/10.1108/JFM-01-2024-0006

Oust A., Westgaard S., Waage J. E., Yemane N. K. (2023), Assessing the explanatory power of dwelling condition in automated valuation models. Journal of Real Estate Research, 1-27. DOI: https://doi.org/10.1080/08965803.2023.2280280

Palo H. K., Sahoo S., Subudhi A. K. (2021), Dimensionality reduction techniques: Principles, benefits, and limitations. Data Analytics in Bioinformatics: A Machine Learning Perspective, 77-107. DOI: https://doi.org/10.1002/9781119785620.ch4

Potrawa T., Tetereva A. (2022), How much is the view from the window worth? Machine learning-driven hedonic pricing model of the real estate market. Journal of Business Research, 144, 50-65. DOI: https://doi.org/10.1016/j.jbusres.2022.01.027

Rico-Juan J. R., De La Paz P. T. (2021), Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. Expert Systems with Applications, 171, 114590. DOI: https://doi.org/10.1016/j.eswa.2021.114590

Sharma H., Harsora H., Ogunleye B. (2024), An optimal house price prediction algorithm: XGBoost. Analytics, 3, 30-45. DOI: https://doi.org/10.3390/analytics3010003

Sibindi R., Mwangi R. W., Waititu A. G. (2023), A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports, 5, e12599. DOI: https://doi.org/10.1002/eng2.12599

Sisman S., Aydinoglu A. C. (2022), Improving performance of mass real estate valuation through application of the dataset optimization and Spatially Constrained Multivariate Clustering Analysis. Land Use Policy, 119, 106167. DOI: https://doi.org/10.1016/j.landusepol.2022.106167

Soltani A., Heydari M., Aghaei F., Pettit C. J. (2022), Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. Cities, 131, 103941. DOI: https://doi.org/10.1016/j.cities.2022.103941

Song Y., Ma X. (2024), Exploration of intelligent housing price forecasting based on the anchoring effect. Neural Computing and Applications, 36, 2201-2214. DOI: https://doi.org/10.1007/s00521-023-08823-3

Stang M., Krämer B., Nagl C., Schäfers W. (2023), From human business to machine learning—methods for automating real estate appraisals and their practical implications. Zeitschrift Für Immobilienökonomie, 9, 81-108. DOI: https://doi.org/10.1365/s41056-022-00063-1

Steurer M., Hill R. J., Pfeifer N. (2021), Metrics for evaluating the performance of machine learning based automated valuation models. Journal of Property Research, 38, 99-129. DOI: https://doi.org/10.1080/09599916.2020.1858937

Theng D., Bhoyar K. K. (2024), Feature selection techniques for machine learning: a survey of more than two decades of research. Knowledge and Information Systems, 66, 1575-1637. DOI: https://doi.org/10.1007/s10115-023-02010-5

Wan Z., Xu Y., Šavija B. (2021), On the use of machine learning models for prediction of compressive strength of concrete: influence of dimensionality reduction on the model performance. Materials, 14, 713. DOI: https://doi.org/10.3390/ma14040713

Xu X., Zhang Y. (2021), House price forecasting with neural networks. Intelligent Systems with Applications, 12, 200052. DOI: https://doi.org/10.1016/j.iswa.2021.200052

Yang S., Yang X., Wang X. (2024), Estimation and Simultaneous Confidence Bands for Fixed-Effects Panel Data Partially Linear Models. Mathematics (2227-7390), 12. DOI: https:// 10.3390/math12233774

Yang X., El-Fallah G. M., Tao Q., Fu J., Leng C., Shepherd J., Dong H. (2023), Dimensionality reduction for machine learning using statistical methods: a case study on predicting mechanical properties of steels. Materials Today Communications, 34, 105162. DOI: https://doi.org/10.1016/j.mtcomm.2022.105162

Zaki J., Nayyar A., Dalal S., Ali Z. H. (2022), House price prediction using hedonic pricing model and machine learning techniques. Concurrency and computation: practice and experience, 34, e7342. DOI: https://doi.org/10.1002/cpe.7342

Zhan C., Liu Y., Wu Z., Zhao M., Chow T. W. (2023), A hybrid machine learning framework for forecasting house price. Expert Systems with Applications, 233, 120981. DOI: https://doi.org/10.1016/j.eswa.2023.120981

Zhang Y., Rahman A., Miller E. (2023), Longitudinal modelling of housing prices with machine learning and temporal regression. International Journal of Housing Markets and Analysis, 16, 693-715. DOI: https://doi.org/10.1108/IJHMA-02-2022-0033

Zippo V., Robotti E., Maestri D., Fossati P., Valenza D., Maggi S., Papallo G., Belay M. H., Cerruti S., Porcu G. (2025), Development of a Self-Updating System for the Prediction of Steel Mechanical Properties in a Steel Company by Machine Learning Procedures. Technologies, 13, 75. DOI: https:// 10.3390/technologies13020075

Zouhri H., Idri A., Hakkoum H. (2024), Assessing the effectiveness of dimensionality reduction on the interpretability of opaque machine learning-based attack detection systems. Computers and Electrical Engineering, 120, 109627. DOI: https://doi.org/10.1016/j.compeleceng.2024.109627