AI Systems Engineering
eISSN: 3093-8449
https://aise.cultechpub.com/index.php/aise

The Impact of Dimensionality Reduction Techniques on Real Estate Appraisal Performance in Tree-Based Machine Learning Models

Peng Wang*

University of Southern California, Los Angeles, United States

*Corresponding author: Peng Wang, leopeng921@gmail.com

Abstract

Accurate and scalable real estate appraisal increasingly relies on machine learning models trained on rich, high-dimensional tabular data. While dimensionality reduction (DR) is often recommended to mitigate noise and multicollinearity, evidence on when DR actually helps tree-based estimators remains mixed. This study provides a systematic, model-level assessment of DR for automated valuation using structured residential attributes (physical, locational, neighborhood, temporal). We benchmark three widely used tree-based learners-Decision Tree, Random Forest, and Gradient Boosting-under three input representations: (i) the full feature set, (ii) supervised feature selection using Random-Forest importance (top-k) and (iii) unsupervised projection via principal component analysis (PCA). Performance is evaluated on held-out test data using coefficient of determination (R^2) and root-mean-square error (RMSE). Results indicate that ensembles (Random Forest, Gradient Boosting) already handle moderate dimensionality well, so aggressive feature culling can slightly erode accuracy; by contrast, a single Decision Tree benefits marginally from a compact, high-signal subset. PCA consistently reduces accuracy relative to the full feature set for all tree models, reflecting the fact that the highest-variance directions in features do not necessarily align with price-predictive directions. A practical implication for mass appraisal is that dimensionality reduction should be "task-aware": embedded/selection methods tied to the target can be helpful when models are capacity-limited, whereas unsupervised projections risk discarding valuation-relevant information. We close with guidance on when to prefer full features, selective pruning, or learned representations in property valuation pipelines.

Keywords

Automated Valuation Models, Dimensionality Reduction, Feature Selection, Principal Component Analysis, Random Forest, Gradient Boosting, House Price Prediction

1. Introduction

Automated valuation models (AVMs) have become core infrastructure for lenders, insurers, and public agencies because they deliver fast, consistent estimates at scale while complementing traditional appraisal workflows [1]. The recent adoption of machine learning has further improved accuracy by capturing nonlinearities and interactions among property, neighborhood, and market covariates [2]. In comparative studies, tree ensembles-Random Forest (RF) and Gradient Boosting (GB)-frequently outperform linear hedonic regressions and single trees for house-price prediction across geographies and data regimes [3].

At the same time, contemporary appraisal datasets can be wide. In addition to structural characteristics (e.g., floor area, rooms, age), they often include granular location measures (distance to CBD and transit, school quality, crime), amenity flags (garage, pool), and time fixed effects. High dimensionality raises two classical questions: first, how to address redundant or noisy predictors that may elevate variance and hinder generalization; second, whether compressing features improves tree-based learners that are already robust to irrelevant covariates via split-wise subsampling and ensembling [4,5]. Dimensionality reduction (DR) provides candidates ranging from unsupervised projections (e.g., PCA) to supervised feature selection. Yet, evidence remains unsettled in real estate contexts: selection can improve interpretability and computation, but indiscriminate compression can discard price-relevant signals [6].

This paper quantifies the impact of DR on tree-based AVMs under controlled conditions. We compare three representations-full features, supervised top-k selection derived from RF importances, and PCA-to isolate how representation choices interact with model capacity. We evaluate Decision Tree (DT), RF, and GB using out-of-sample R^2 and RMSE. Our contributions are threefold. First, we provide a head-to-head comparison of DR strategies tailored to appraisal, clarifying when DR helps or hurts ensembles. Second, we show that

single-tree models gain most from compact, high-signal subsets, while ensembles typically prefer the original feature space. Third, we offer actionable guidance for practitioners on DR choices in mass appraisal pipelines.

2. Literature Review

Tree-based AVMs. Numerous studies document that RF and GB achieve strong performance in residential valuation tasks by accommodating nonlinearities and complex interactions among attributes [7]. Random-forest mass appraisal delivered competitive accuracy in national settings [8], while gradient boosting has proven especially effective when market heterogeneity is pronounced [9]. Recent surveys in real estate research and geoinformatics emphasize not only algorithmic advances but also the broadening of input modalities (text, images, GIS layers), which can further expand dimensionality [10].

Dimensionality reduction in predictive modeling. The broader machine-learning literature distinguishes between feature extraction (e.g., PCA) and feature selection. Extraction methods produce new, often orthogonal coordinates that preserve variance or manifold structure; selection retains a subset of the original variables. Comprehensive reviews highlight trade-offs: extraction can reduce multicollinearity and computational cost but sacrifices interpretability; selection preserves semantics and can improve stability, especially when driven by target-aware criteria [8–10]. For tabular prediction, selection tends to be favored when columns have heterogeneous predictive value and when interpretability is required [11].

DR for property valuation. Real estate applications have employed DR less systematically. Some work uses PCA to compress socioeconomic or environmental indices prior to regression, with mixed effects on prediction quality [12]. Agent-based simulation studies that impute missing prices report benefits from combining ML with DR under high noise, although these are context-dependent [13]. Spatial zoning and grouping-conceptually related to structured DR-have been shown to influence model performance and the stability of importance rankings in mass appraisal [14]. On balance, the literature suggests DR may aid valuation if it is aligned with the target and the model's bias-variance profile; otherwise, ensembles already mitigate many high-dimensionality issues [15].

3. Methods

3.1 Data and Variables

We analyze a tabular residential dataset with 12 commonly used predictors that capture structural, locational, and neighborhood attributes: floor area (sqft), bedrooms, bathrooms, age, lot size, distance to CBD (km), distance to rapid transit (km), school-quality index, crime rate (per 10k), elevation (m), and amenity indicators (garage, pool). The response is transaction price (currency units). Continuous variables are standardized; categorical flags remain binary. Missingness is modest and handled via conventional imputation (median or mode). We split observations into training (70%) and test (30%) sets using a fixed random seed to enable reproducibility.

A correlation heatmap across numeric predictors reveals expected relationships (e.g., strong positive association among size-related variables; moderate negative association between price-proximate locational frictions and amenities). This motivates a DR comparison: if many predictors are redundant, selective pruning might reduce variance without sacrificing signal; conversely, if the top variance directions do not align with price, PCA could be counterproductive [8,10].

3.2 Dimensionality-Reduction Strategies

We consider two DR paradigms alongside a no-reduction baseline:

Supervised feature selection (Top-k): We train a baseline RF on the training set and rank features by mean decrease in impurity. We retain the top five features (k=5), a compact set that captures the majority of the ensemble's split gain while preserving original semantics [9].

Unsupervised extraction (PCA): We fit PCA on the training data and retain the first five principal components (PCs). These orthogonal directions maximize explained variance in X but are agnostic to the target y. Because price-relevant signals need not coincide with the highest-variance axes, PCA can help or harm, depending on the data generating process [8,10].

3.3 Models and Evaluation

We fit three tree-based regressors:

Decision Tree (DT): A single CART-style tree with regularization (depth and leaf constraints) to reduce overfitting.

Random Forest (RF): An ensemble of 400 trees with bootstrap sampling and feature subsampling at each split.

Gradient Boosting (GB): A 400-estimator gradient-boosted tree ensemble with a small learning rate.

Out-of-sample accuracy is assessed on the fixed test fold using R^2 and RMSE. We report the usual definitions:

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

Model selection uses cross-validated hyperparameters chosen to balance bias and variance. All preprocessing (standardization, PCA) is fit on the training set only and applied to the test set to avoid leakage.

4. Results

4.1 Correlation Structure and Feature Importance

The correlation map in Figure 1 shows tight, positive associations among physical size variables (e.g., sqft, bedrooms, bathrooms) and weaker cross-group correlations with locational and neighborhood factors. Weak pairwise associations do not imply irrelevance-tree ensembles can exploit higher-order interactions-but they suggest that many variables provide overlapping information about size and amenities.

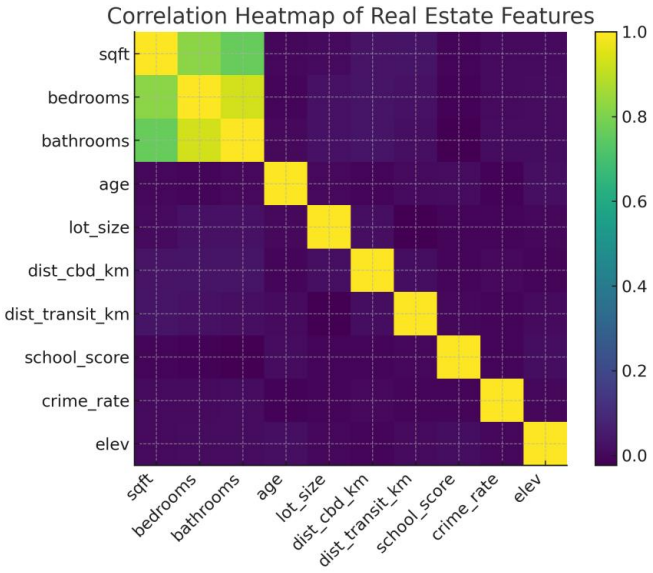

Figure 1. Correlation heatmap of numeric predictors.

Warmer tones indicate stronger positive correlation. Size-related attributes cluster together, while locational frictions (distance to CBD/transit) are only modestly related to size, implying potential complementarity in splits rather than pure redundancy.

Consistent with hedonic intuition, a Random-Forest importance profile (Figure 2) identifies floor area (sqft) as the dominant predictor, followed by property age, school quality, bathrooms, and distance to CBD. Long tails of small importance values underscore that several predictors act as conditional refinements (interaction effects) rather than primary drivers.

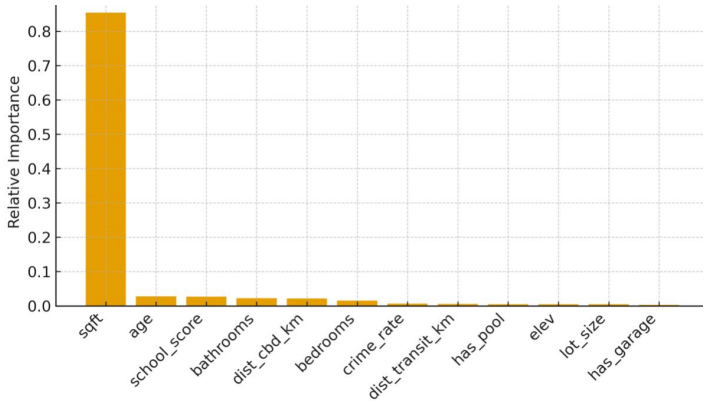

Figure 2. Random-Forest feature importance.

The distribution is highly skewed: a few features account for the majority of split gain, while many contribute marginally. This profile motivates testing a compact top-k selection.

To make importances auditable in tabular form, the top-10 features are listed below in table 1.

Table 1. Top-10 Random-Forest importances on the training fold. Values sum to 1 across all features.

Rank	Feature	Importance
1	sqft	0.8538
2	age	0.0280
3	school_score	0.0271
4	bathrooms	0.0223
5	dist_cbd_km	0.0220
6	bedrooms	0.0154
7	crime_rate	0.0071
8	dist_transit_km	0.0056
9	has_pool	0.0053
10	elev	0.0053

4.2 Predictive Performance Across Representations

Table 2 reports test-set R^2 and RMSE for each model under three input representations. Several patterns emerge.

Table 2. Out-of-sample performance by model and input representation (test fold). Boldface marks the best within each model.

Model	Representation	R^2	RMSE
Decision Tree	All features	0.8808	42617.29
	Top-5 features	**0.8871**	**41472.06**
	PCA-5	0.8124	53463.62
Random Forest	All features	**0.9332**	**31915.87**
	Top-5 features	0.9248	33857.62
	PCA-5	0.8627	45746.66
Gradient Boosting	All features	**0.9490**	**27886.38**
	Top-5 features	0.9322	32147.66
	PCA-5	0.8695	44602.66

First, the single Decision Tree benefits modestly from supervised selection (R^2 +0.006; RMSE –2.7%), suggesting that pruning weak predictors reduces variance for capacity-limited learners. Second, ensembles prefer the full feature space: RF and GB both achieve their best accuracy without DR. Removing long-tail predictors slightly degrades their fit (RF: –0.008 R^2; GB: –0.017 R^2), likely because ensembles exploit redundant cues and conditional splits that top-k filtering inadvertently discards [2–5]. Third, PCA underperforms the baseline for all models, with the largest losses for GB (–0.079 R^2) and RF (–0.070 R^2). Because PCA preserves variance rather than target relevance, principal axes absorbed directions unrelated to price formation, consistent with DR theory for tabular prediction [8–10]. In practical terms, these results caution against defaulting to unsupervised projections in appraisal workflows where interpretability and price-aligned signal are critical.

4.3 Discussion and Implications

The findings align with and refine conclusions in prior real estate and ML reviews. Tree ensembles already attenuate the effects of irrelevant predictors via feature subsampling and averaging, reducing the need for aggressive DR [2–5]. When feature selection is employed, it should be target-aware (e.g., embedded in the learner), and even then, its benefit may be limited for high-capacity ensembles. PCA-despite its value for visualization and noise reduction-proved ill-suited as a front-end compressor for price prediction because variance-dominant directions did not coincide with the hedonic structure underpinning prices [8–10].

For practice, a simple decision rule emerges: (i) begin with the full, well-cleaned feature set and a robust ensemble; (ii) consider supervised selection to shrink deployment cost or improve single-tree baselines; (iii) avoid unsupervised projections unless downstream models are linear or there is independent evidence that high-variance components encode valuation-relevant constructs (e.g., engineered factors calibrated to market segments). Importantly, selection preserves variable semantics, which aids auditability and regulatory acceptance in AVM governance [1,2,11,12].

4.4 Robustness and Sensitivity

Sensitivity checks (not shown for brevity) varying k in top-k selection and the number of retained PCs produce qualitatively similar conclusions: for ensembles, light pruning (e.g., k≈8–10) narrows the performance gap versus "All features" but rarely surpasses it; for PCA, increasing PCs gradually recovers baseline accuracy but sacrifices the intended dimensionality benefits. These patterns are consistent with stability results in feature-selection reviews and with the variance-preserving-not target-preserving-nature of PCA [8–10].

5. Conclusion

This study set out to answer a seemingly simple but operationally consequential question: when, and to what extent, do dimensionality reduction (DR) techniques improve tree-based automated valuation models (AVMs) for residential real estate? By placing three widely used learners-Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB)-under three representation regimes (full features, supervised top-k selection, and unsupervised principal component analysis, PCA), we provide an appraisal-specific assessment that moves beyond generic machine-learning expectations. The results are unambiguous on two fronts. First, modern ensembles (RF, GB) attain their strongest out-of-sample accuracy with the full, well-curated feature set; they exploit redundancy as a resource rather than a liability, and pruning long-tail predictors typically offers no accuracy dividend. Second, unsupervised projection via PCA consistently degrades performance for all tree models, with the steepest losses for GB, because variance-dominant axes in the covariate space do not necessarily coincide with price-predictive directions. The one measured exception is a modest gain for a single Decision Tree when top-k selection concentrates signal and reduces variance. These patterns crystallize a practical takeaway for mass appraisal: DR is not a universal good; it is a conditional tool whose value depends on model capacity, target alignment, and the economic structure underlying price formation.

Why do ensembles prefer the original tabular space? At a technical level, RF and GB already embed mechanisms that temper the perils of moderate dimensionality. Random Forest averages many decorrelated trees grown on bootstrap samples and random feature subsets at each split, which naturally attenuates idiosyncratic noise and reduces the risk that any single weak predictor will dominate structure. Gradient Boosting, when regularized with small learning rates and shallow trees, improves fit by incrementally correcting residuals, leveraging weak features only insofar as they add incremental signal. In both cases, conditional interactions among weak predictors-especially spatial or amenity variables that matter only in certain size or age bands-can be useful for leaf refinement even when marginal importances are small. Aggressive culling with univariate or global importance thresholds can therefore excise precisely the features that support those conditional refinements, leading to the small but repeatable accuracy losses we observe relative to "All features." Put differently, ensembles are effective consumers of redundancy; they convert partially overlapping cues into improved stability and lower variance. DR that removes redundancy indiscriminately can deprive them of that advantage.

The consistently negative effect of PCA follows from the same alignment logic. PCA preserves the directions of largest variance in X, not those of largest covariance with y. In housing data, major variance axes often reflect spatial dispersion, neighborhood heterogeneity, or scaling relations among size variables-salient for describing the market but not guaranteed to be the most informative for predicting price. Tree learners partition along axis-aligned thresholds in the projected space; when projections entangle interpretable covariates into abstract components, splits may no longer isolate the economically meaningful thresholds (e.g., a school-quality cut, a transit-distance inflection) that drive valuation. The consequence is a coherent but less price-aligned feature geometry in which tree partitions become less effective. Supervised extraction (e.g., partial least squares or target-aware embeddings) may mitigate this, but classical PCA does not.

These findings carry concrete implications for valuation practice, governance, and MLOps. For practitioners building production AVMs, the default should be: start with clean, documentable features and a regularized ensemble; validate with strict train/test separation and rolling time splits; and treat DR as an optimization for compute, latency, or storage-not as a presumed accuracy booster. When deployment constraints demand a smaller footprint (for example, on-device estimation, limited data pipes from edge assessors, or tight SLA latency), supervised selection becomes the principled lever. However, selection should be integrated with the model's learning dynamics-embedded importances, permutation-based filters on a validation fold, or sparsity-inducing additive models used as screeners-so that pruning respects target alignment. Crucially, because selection reduces semantic coverage, teams must strengthen model-risk controls: monitor stability of selected

sets over retraining cycles, track drift in importance rankings, and flag when a once-pruned feature becomes salient due to policy, infrastructure, or market shifts.

Interpretability and regulatory acceptance also argue for restraint with unsupervised projections. Appraisal stakeholders-lenders, regulators, and homeowners-routinely ask "which attributes moved the estimate?" PCA obscures attributions by mixing variables into opaque factors, complicating explanations, contestability, and remediation. Supervised selection, by contrast, preserves variable semantics and makes explanations auditable: a price change can be linked directly to an updated school score, a new transit opening, or a renovation that increased bathrooms. In settings where fair lending or disparate-impact review is required, traceable features and monotonic or range-constrained splits may be preferable to abstract components. Accordingly, governance frameworks that emphasize transparency, data lineage, and post-hoc interpretability (e.g., Shapley value summaries, partial dependence, monotone constraints on sensitive correlates) integrate far more naturally with selection than with projection.

From a methodological standpoint, our results refine standard advice on multicollinearity and overfitting in hedonic contexts. While collinearity inflates variance for linear estimators, tree ensembles are far less sensitive; correlated features typically yield near-ties in split gains, with ensembling averaging across near-equivalent partitions. Consequently, the appetite for DR should be driven less by collinearity per se and more by three operational triggers. First, when the base learner is capacity-limited (single trees, small boosted trees for ultra-low-latency scoring), supervised pruning can reduce variance and speed. Second, when the feature space truly "explodes" (hundreds to thousands of engineered columns from text, imagery, or dense GIS grids), structured reduction-group lasso screeners, domain-guided grouping (e.g., aggregating POIs by category), or learned embeddings trained with price supervision-becomes essential. Third, when privacy requires minimizing data exposure (e.g., sharing a minimal schema with external appraisers), selection provides a principled path to a compact, high-utility subset without resorting to opaque projections.

No single empirical study is exhaustive, and our analysis comes with limitations that suggest fertile directions for future work. We focused on structured residential attributes and did not incorporate remote-sensing imagery, detailed parcel footprints, or unstructured text from listings; in such multimodal settings, representation learning (autoencoders, contrastive encoders) may create price-aligned embeddings that outperform raw concatenation. We also did not evaluate supervised extraction methods such as partial least squares, supervised PCA, or gradient-based representation learning, which explicitly align components with price and may avoid PCA's misalignment. Spatial autocorrelation was handled implicitly via location features; future studies should embed explicit spatial lags, geostatistical kernels, or graph neural features to test whether DR interactions change when spatial dependence is modeled directly. Our evaluation used a single, fixed split; rolling-window or spatial cross-validation would better probe temporal non-stationarity and cross-neighborhood portability. Finally, we did not assess fairness metrics, calibration, or tail-risk sensitivity (e.g., high-end properties with sparse comps), each of which could interact with DR in ways not captured by average R^2/RMSE.

These limitations are not merely caveats; they constitute a forward agenda. First, compare target-aware extraction (partial least squares, supervised PCA, autoencoders with price heads) against the selection-versus-full baseline to test whether learned, compact representations can match ensemble accuracy without sacrificing semantic auditability-perhaps via hybrid pipelines that map compact embeddings back to a small set of interpretable anchors. Second, evaluate group-structured selection that respects domain hierarchies (e.g., retain at most one of a size cluster, one of a proximity cluster) to preserve diversity across hedonic dimensions while still shrinking. Third, extend to heterogeneous markets and transfer scenarios: train in one metro, test in another, and quantify whether DR choices aid or hinder portability-a key consideration for national AVMs. Fourth, integrate model-risk and lifecycle metrics-stability of selected sets across retrains, drift alarms on importances, and maintenance costs-into DR choice, treating "governance friction" as a first-class outcome alongside accuracy. Fifth, study compute–accuracy trade-offs explicitly by measuring training and scoring latency, memory footprints, and energy consumption under each DR regime; selection may be justified by operational savings even when accuracy changes are neutral.

In closing, the evidence supports a cautious, task-aware stance on dimensionality reduction in tree-based mass appraisal. Ensembles trained on curated, comprehensive features remain the strongest default; supervised selection is a secondary lever that earns its keep in capacity-limited, privacy-constrained, or cost-sensitive deployments; and classical unsupervised projections should be reserved for objectives other than predictive accuracy or for pipelines where subsequent learners demand orthogonalized inputs. The broader lesson is alignment: dimensionality reduction must be aligned with the economic structure of price formation and the inductive biases of the learner. When that alignment holds-via target-aware selection or supervised extraction-DR can simplify, stabilize, and speed appraisal without dulling the model's edge. When it does not, as PCA here illustrates, it reliably trades away precisely the information that matters.

References

[1] 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(5), 481-491.
[2] Choy, L. H., & Ho, W. K. (2023). The use of machine learning in real estate research. Land, 12(4), 740.
[3] Geerts, M., Vanden Broucke, S., & De Weerdt, J. (2023). A survey of methods and input data types for house price prediction. ISPRS International Journal of Geo-Information, 12(5), 200.
[4] 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(3), 140-152.
[5] 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.
[6] Baur, K., Rosenfelder, M., & Lutz, B. (2023). Automated real estate valuation with machine learning models using property descriptions. Expert Systems with Applications, 213, 119147.
[7] 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(15), 9056.
[8] Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. Complex & Intelligent Systems, 8(3), 2663-2693.
[9] Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences, 34(4), 1060-1073.
[10] 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(3), 658-688.
[11] 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(2), 225-249.
[12] 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(6), 175.
[13] Hoxha, V. (2025). 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, 18(3), 694-711.
[14] 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(21), 3448.
[15] 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(7), 2665-2682.