ZAMBIA INFORMATION COMMUNICATION TECHNOLOGY (ICT) JOURNAL

Volume 8 (Issue 1) (2024) Pages 34-46

Utilizing Machine Learning for Accurate Property Valuation: A Regression Model Analysis

Simone Chishala Kaoma
School of Engineering and Technology
Mulungushi University
Kabwe, Zambia
202201098@mu.edu.zm

Brian Halubanza
School of Engineering and Technology
Mulungushi University
Kabwe, Zambia
bhalubanza@mu.edu.zm

***Abstract*— This research investigated the application of machine learning (ML) regression models to improve property valuation accuracy, addressing limitations of traditional methods. The study applied Random Forest (RF) and Support Vector Regression (SVR) models to a dataset of 59,180 property records from the Luanshya Municipal Council. Key features such as LAND_VALUE, MARKET_VALUE, and IMPROVEMENT_VALUE were used as inputs. The models' performance was evaluated using the Data, Reasoning, and Usefulness (DRU) Evaluation Framework. Results showed that both RF and SVR outperformed traditional methods, with RF achieving an R^2 of 0.9995. Machine Learning models demonstrated potential for more accurate property valuations, enabling fairer tax assessments, reduced manual effort, and improved urban planning decisions. Future research should address data quality and model explainability challenges.**

Keywords—Property Valuation, Machine Learning, Random Forest, Support Vector Machines, Data Quality, Explainability.

I. Introduction

Accurate property valuation is critical for municipal governance, influencing property taxes and contributing to financial stability. However, the dynamic nature of real estate markets, driven by economic trends, government policies, and the varying characteristics of properties, presents challenges for traditional valuation methods. These methods often fail to adapt to market fluctuations, leading to inaccuracies that impact tax assessments and urban planning decisions [1, 2].

To address these challenges, this study utilizes machine learning algorithms, particularly regression-based models, to enhance the accuracy and efficiency of property valuations. Machine learning techniques, which can analyze large datasets of real estate variables, offer the ability to adapt over time and better reflect subtle market trends. By leveraging key property features such as LAND_VALUE, MARKET_VALUE, and IMPROVEMENT_VALUE, machine learning models provide a more robust and adaptable framework for property valuation [3, 4].

The performance of these machine learning models is assessed using the Data, Reasoning, and Usefulness (DRU) Evaluation Framework [5]. This framework integrates insights from the Technology Acceptance Model (TAM) [6], the Task Technology Fit (TTF) model [7], and data quality dimensions [8], offering a comprehensive evaluation of data quality, model reasoning, and practical utility. By applying the DRU framework, this study ensures that the machine learning models used for property valuation not only offer high predictive accuracy but also align with established real estate principles and provide tangible value to municipal valuation processes. The objectives of this study are twofold: to assess the accuracy and interpretability of machine learning models in predicting property values using key variables**,** and to ensure that the model predictions align with established real estate principles, identifying the most influential variables in property valuation**.**

II. LITERATURE REVIEW

Real estate valuation plays a vital role in the global economy, influencing various stakeholders, including buyers, sellers, developers, property owners, renters, and governments. Valuation determines a property's market value, which is influenced by factors such as property features, location, market dynamics, and individual preferences [9]]. Traditionally, property valuation methods have focused on three primary approaches:

- Sales Comparison Approach: This method compares the subject property to recently sold properties, factoring in attributes like the number of bedrooms, bathrooms, lot size, and square footage [10]. It aligns with the hedonic regression model, where buyers assign different weights to various property characteristics [11].
- Income Approach: Focused on income-generating properties, this approach calculates the present value of future cash flows expected from the property [12].
- Cost Approach: This method estimates the value based on the cost to replace the structure, adding the value of the land and subtracting depreciation [13]. It is typically used for new constructions.

These traditional methods blend mathematical tools and subjective adjustments by human appraisers, which can lead to variations in estimated values [14].

Recent technological advancements have enabled the use of mass appraisals, which assess groups of properties using standardized data and methods. This technique is often used for property tax assessments and development planning [15]. However, traditional statistical methods and large datasets often struggle to account for locational differences, limiting the effectiveness of mass appraisals as the sole valuation method [16]. Machine learning (ML) offers a promising alternative, capable of analyzing large datasets and identifying complex, nonlinear relationships within property valuation [17].

Despite its potential, the adoption of ML in real estate valuation faces several challenges:

- Model Specification and Parameterization: The selection of appropriate parameters in ML models, such as the number of hidden layers and nodes in artificial neural networks (ANNs), significantly impacts performance and can result in varying outcomes [18].
- Inconsistent Results and Measurement Error: Different runs of the same model may produce inconsistent results due to random initialization of weights in ANNs [19].
- Computational Time: The complexity and size of data can result in long computational times, potentially deterring ML adoption [20].
- Lack of Transparency: Some ML models operate as "black boxes," making it difficult for practitioners to understand how valuations are derived, which can hinder trust and acceptance [21].

The DRU framework (Data, Reasoning, and Usefulness) provides a valuable lens for assessing the effectiveness of ML in property valuation. This framework highlights three critical factors:

- Data Quality: High-quality, relevant, accurate, and complete data is essential for reliable valuations [22].
- Reasoning, Alignment, and Explication: The chosen ML technique should align with the valuation task’s principles and deliver explainable results [23].
- Usefulness and Consistency of Results: ML models should provide consistent, reliable outputs to improve accuracy and efficiency in valuation [24].

The DRU framework suggests that ideal ML models should:

- Be adaptable to various data sources and scenarios [25].
- Effectively account for the location of the property [26].
- Have low barriers to entry in terms of ease of use and adoption [27].
- Deliver performance improvements that justify computational costs [28].

While no ML technique currently satisfies all these criteria, research continues to advance towards more effective solutions [29].

Machine learning models have demonstrated the potential to significantly improve the accuracy of real estate valuations compared to traditional approaches such as multiple regression analysis (MRA) [18]. The review of literature published since 2000 indicates that as computing power and data availability have improved, ML techniques such as ANNs, regression trees, and support vector machines (SVM) have shown varying degrees of success in property valuation [30]. However, careful attention to data quality, model transparency, and computational costs is essential for successful implementation [21].

Strengths of ML Models

- Capturing Nonlinearities: ML models, especially regression trees like random forest and gradient boosting, excel at capturing complex, nonlinear relationships between property characteristics and values [17]. These relationships, often tied to location and time, are frequently missed by traditional hedonic regression models [17].
- Incorporating Locational Data: ML models can effectively incorporate locational variables (e.g., proximity to amenities, crime rates, demographic data), enhancing the accuracy of valuations compared to traditional approaches [31].
- Improved Accuracy: Studies show that ML models like boosted regression trees consistently outperform traditional regression models, KNN, SVM, and various ANN approaches [19].

Challenges of ML Models

- Data Quality and Model Specification: The performance of ML models heavily depends on the size and quality of the datasets. Smaller or poorly specified datasets can lead to overfitting and limit the generalizability of the results [22].
- Transparency and Ease of Use: Complex ML models, such as ANNs and hybrid approaches, may lack transparency, making it difficult for users to understand the decision-making process, which can hinder widespread adoption [23].
- Computational Costs: Some advanced ML techniques, such as boosting or bagging, require significant computational resources, potentially making them less practical for real-world applications [20].

To maximize the benefits of ML in real estate valuation, the DRU framework emphasizes the need to evaluate:

- Data Quality: Ensuring access to high-quality, relevant, and accurate data is crucial for effective ML model performance [24].
- Reasoning and Transparency: The alignment and explainability of the model should be prioritized to build trust among practitioners [23].
- Usefulness and Consistency: ML models should provide consistent, reliable results that improve the efficiency and accuracy of valuations [24].

Finally, demonstrating the economic benefit of ML adoption, particularly by translating small improvements in performance into tangible financial gains, is critical for encouraging the use of ML models in property valuation [28].

III. METHODS

This study utilized the Data, Reasoning, and Usefulness (DRU) Evaluation Framework, introduced by [5], to assess the performance of machine learning-based property valuation systems. The DRU framework synthesizes elements from the Technology Acceptance Model (TAM) [6], Task Technology Fit (TTF) model [7], and data quality dimensions [8], while integrating practical insights from real estate professionals. The framework categorizes evaluation criteria into three core components of an information system: input, process, and output. This structured approach ensures a comprehensive evaluation of data quality, model reasoning, and the practical utility of the system [5]. Table 1 provides a summary of the criteria outlined by [5] within the DRU framework.

TABLE I. DRU FRAMEWORK

Category	***Criteria***	***Theoretical Basis***
Input	Data quality, Ease of Use	TAM: Perceived ease of use; Data quality dimensions: Intrinsic, contextual, accessibility, representational [1]
Process	Reasoning, Explication	TTF: Alignment between technology and task; Explicit consideration of locational elements [1].
Ouput	Usefulness, Result Consistency	TAM: Perceived usefulness; Professional principles: Statistical consistency and transparency [1].

A. Data Collection

The dataset, comprising 59,180 property valuation records, was sourced from the Luanshya Municipal Council database. To ensure its suitability for analysis, the dataset was processed in accordance with the DRU framework's input criteria, with a focus on two key aspects:

1. **Data quality**: The dataset was selected for its high intrinsic, accessibility, contextual, and representational quality, ensuring that it was accurate, complete, and relevant for the analysis.
2. **Ease of use**: The data underwent rigorous preprocessing to enhance its usability, including formatting and cleaning tasks to eliminate errors, inconsistencies, and redundancies. This step was crucial in streamlining the modeling process and preparing the dataset for subsequent analysis.

The dataset, sourced from the Luanshya Municipal Council database, comprised of 59,180 property valuation records. It was processed according to the DRU framework's input criteria, focusing on data quality and ease of use.

B. Exploratory Data Analysis

To satisfy the DRU framework's reasoning and alignment criteria, a comprehensive Exploratory Data Analysis (EDA) was conducted on the dataset. This involved applying a range of EDA techniques to examine the data's underlying structure, relationships, and patterns, thereby ensuring that the analysis was grounded in a thorough understanding of the data.

1. **Summary Statistics**: Measures of central tendency (mean, median) and dispersion (standard deviation, range) were calculated to provide insights into the dataset's distribution and identify potential data issues.
2. **Visualization**: Histograms were plotted for each variable to visually assess the distribution and skewness. Skewed variables were identified as candidates for transformation to improve model performance.
3. **Correlation Analysis**: A correlation matrix was generated to identify relationships between variables, using Pearson's correlation coefficient r, defined as:

$$R= \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}} \tag{1}$$

C. Data Preprocessing

To align with the input and processing requirements of the DRU framework, a comprehensive data preprocessing procedure was implemented. The following steps were undertaken to guarantee data quality, transparency, and adherence to the framework's specifications:

1. **Handling Missing Values**
   Continuous variables were imputed using the median, while categorical variables were imputed with the mode. A missing values matrix was created to visualize gaps in the data. Little's MCAR test was applied to assess the randomness of missing data.
2. **Outlier Detection and Transformation**
   Outliers were identified using the Interquartile Range (IQR) method. Extreme values were log-transformed to reduce their influence on model outcomes and ensure data consistency.
3. **Transparency**

All preprocessing steps were carefully documented to ensure reproducibility and compliance with best practices, ensuring the methodology was clear and aligned with the DRU framework.

D. Model Development

Model development was guided by the reasoning and alignment criteria of the DRU framework. Five machine learning models were selected for implementation: Linear Regression, Decision Trees, Random Forest, Gradient Boosting, and Support Vector Machines (SVM). These algorithms were chosen based on their theoretical relevance and proven effectiveness in property valuation tasks. To ensure explain ability and transparency, Decision Trees and Random Forests were prioritized, as they provide interpretable outputs that support reasoning and facilitate understanding of model decisions. Consistency across models was maintained by adhering to a standardized training and evaluation process, ensuring reliable and comparable results.

Table II outlines the models employed in this study, along with their corresponding mathematical formulations. To evaluate the generalizability of each model, we trained them on the training dataset and validated their performance on the testing dataset. The dataset was randomly partitioned into a training set (80% of the total data) and a testing set (20% of the total data), ensuring a robust assessment of the models' ability to generalize to unseen data. A 10-fold cross-validation was employed to ensure the robustness and reliability of the model evaluations.

TABLE II. MODEL EQUATIONS

Model	***Equation***
Linear Regression	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta n X n + \epsilon$ (2)
Decision Trees	Decision trees recursively partition the feature space into regions. The prediction for a given region is the average (regression) or majority vote (classification) of the target values in that region.
Random Forest	$\mathrm{Y} = \frac{1}{\mathrm{B}} \sum_{\mathrm{b}=1}^{\mathrm{B}} \mathrm{T_b}\ (\mathrm{x}\,;\ominus_{\mathrm{b}})$ (3)
Gradient Boosting	$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$ (4)
Support Vector Machine (SVM)	$\sum_{i=0}^{n} \alpha_i K(x_{i,}\ x) + b$ (5)

.

E. Evaluation

Model performance was evaluated using metrics aligned with the output criteria of the DRU framework:

- **Accuracy Metrics:** Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) were calculated.
- **Result Transparency:** Outputs were examined for consistency and alignment with task requirements.
- **Consistency:** Models were assessed for their ability to improve decision-making quality and productivity in property valuation tasks.

TABLE III. ACCURACY METRICS

Metric	***Equation***
Mean Squared Error (MSE)	$MSE = \frac{1}{n}(\sum_{i=0}^{n}(Y_i - \hat{Y}_i)^2$ (6)
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{MSE}$ (7)
R-squared (R^2)	$R^2 = \frac{\sum_{i=0}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=0}^{n}(Y_i - \bar{Y}_i)^2}$ (8)

IV. FINDINGS

The performance of the machine learning models was evaluated based on several key criteria: model accuracy, explainability, and consistency with real estate valuation principles. These criteria were assessed in relation to the DRU framework, specifically focusing on the input, process, and output components; data quality, reasoning, and usefulness. This section presents the findings in terms of how well the models met these criteria, offering insights into their ability to handle the data, provide interpretable outputs, and deliver useful results aligned with industry standards.

A. Data Collection

Table IV provides a detailed overview of the variables utilized in the analysis, showcasing a balanced mix of quantitative and categorical data types. This deliberate selection of variables ensures a comprehensive examination of the various factors influencing property valuation, enabling a nuanced understanding of the complex relationships between these factors.

TABLE IV. DATA VARIABLES

Variable Name	***Type***	***Description***
TOTAL_VALUE	DEPENDENT	Total rateable value of each property
HECTORAGE	INDEPENDENT	Size of the land in hectares
LAND_VALUE	INDEPENDENT	Base monetary value
SURFACE_AREA	INDEPENDENT	Value attributed to the size of property characteristics
MARKET_VALUE	INDEPENDENT	Market value influenced by external economic conditions
IMPROVEMENT_VALUE	INDEPENDENT	Value added through developmental improvements
CAT	INDEPENDENT	Category of the property (e.g., Residential, Commercial)
AREA	INDEPENDENT	Location where the property is located

The dataset for this study consisted of a range of features that impact the total rateable value (TOTAL_VALUE) of properties. To identify the most influential factors, a selection of key independent variables was made based on their relevance to property valuation. These variables can be categorized into two main groups:

1. Quantitative Variables: HECTARAGE (size of the land in hectares), LAND_VALUE (base monetary value of the land), SURFACE_AREA (size-related value), MARKET_VALUE (value influenced by external economic conditions), and IMPROVEMENT_VALUE (value added through property improvements).
2. Qualitative Variables: Categorical variables such as CAT (property category—e.g., Residential, Commercial) and AREA (location of the property).

B. *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) provided critical insights into the dataset's structure and the relationships between variables. Summary statistics in Table V revealed that most properties had relatively low values, with means and medians significantly below maximums. For instance, the mean LAND_VALUE was 2.15 million compared to a maximum of nearly 100 billion, and TOTAL_VALUE had a mean of 8.03 million versus a maximum of 110 billion. These disparities, driven by a few high-value outliers such as luxury estates, resulted in high standard deviations and right-skewed distributions. Addressing these outliers through transformations or robust methods was essential for accurate property value modeling.

TABLE V. SUMMARY STATISTICS

statistic	***area***	***land_value***	***surface_area***	***market_value***	***improvement_value***	***total_value***
Count	59180	59179	59157	59172	59176	59180
Mean	7.96	2.15E+06	0.728459	4.57E+06	5.86E+06	8.03E+06
Std Dev	2.53	4.10E+08	10.92873	1.05E+07	6.98E+07	4.56E+08
Min	1.00	0.00	0.00	0.00	0.00	0.00
25%	6	10400	0.11745	240740.7	42000	54800
50%	8	14400	0.209375	320987.7	72000	86600
75%	10	41400	0.590625	460000	315000	369000
Max	11	9.97E+10	1745.68	8.17E+08	1.03E+10	1.10E+11

Fig. 1 displays histograms of key variables, revealing the skewed distribution of the data. Notably, the LAND_VALUE histogram exhibits a pronounced skewness, characterized by a cluster of lower values and a long tail of high-value properties. This asymmetry reflects the significant variability in property values within the dataset, highlighting the presence of both affordable and luxury properties..

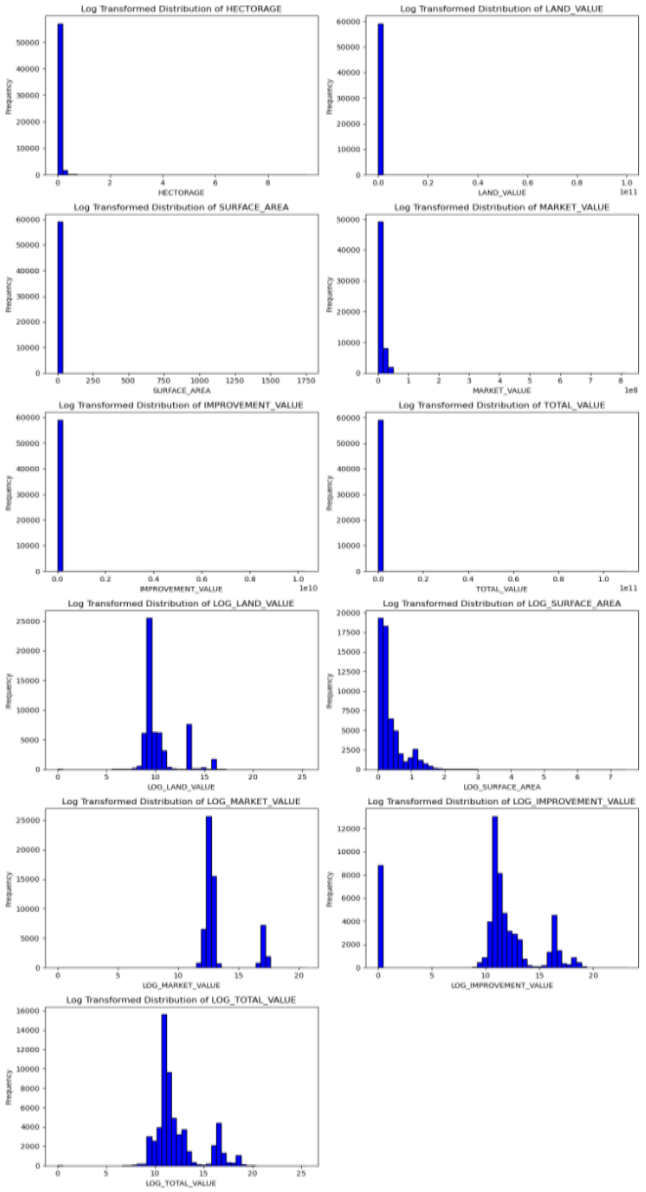
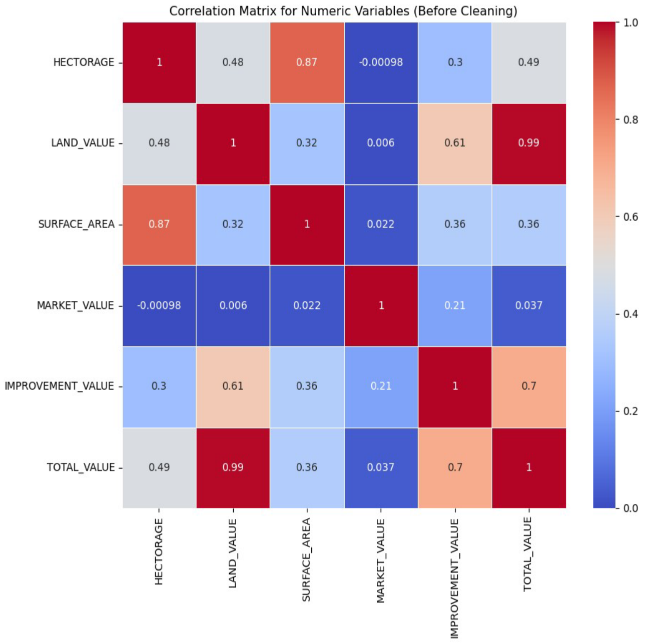


Fig.1. Histograms of key variables

The correlation analysis, as illustrated in Fig. 2, revealed strong relationships between the variables. Notably, the correlation matrix showed a very high correlation coefficient (0.99) between TOTAL_VALUE and LAND_VALUE, indicating that land value is the primary driver of property worth. Furthermore, IMPROVEMENT_VALUE was also found to be strongly correlated (0.70) with TOTAL_VALUE, highlighting the significant impact of property enhancements on the overall value of a property.



	HECTORAGE	LAND_VALUE	SURFACE_AREA	MARKET_VALUE	IMPROVEMENT_VALUE	TOTAL_VALUE
HECTORAGE	1	0.48	0.87	-0.00098	0.3	0.49
LAND_VALUE	0.48	1	0.32	0.006	0.61	0.99
SURFACE_AREA	0.87	0.32	1	0.022	0.36	0.36
MARKET_VALUE	-0.00098	0.006	0.022	1	0.21	0.037
IMPROVEMENT_VALUE	0.3	0.61	0.36	0.21	1	0.7
TOTAL_VALUE	0.49	0.99	0.36	0.037	0.7	1

Fig. 2. Corelation Matrix

C. *Data Preprocessing*

The identification and management of missing values and outliers were essential steps in ensuring the accuracy and reliability of the analysis. To visualize the distribution of missing values, a heatmap was generated, as shown in Fig. 3. This heatmap provides a clear representation of the missing value patterns, allowing for the identification of variables with high rates of missingness and informing the development of effective imputation strategies.

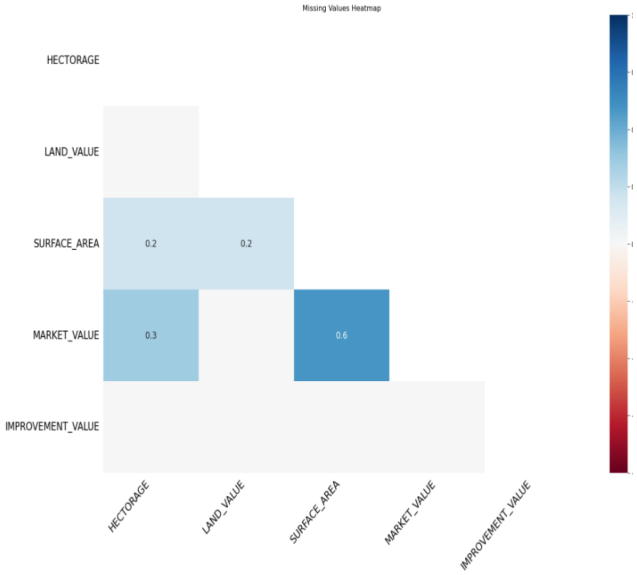
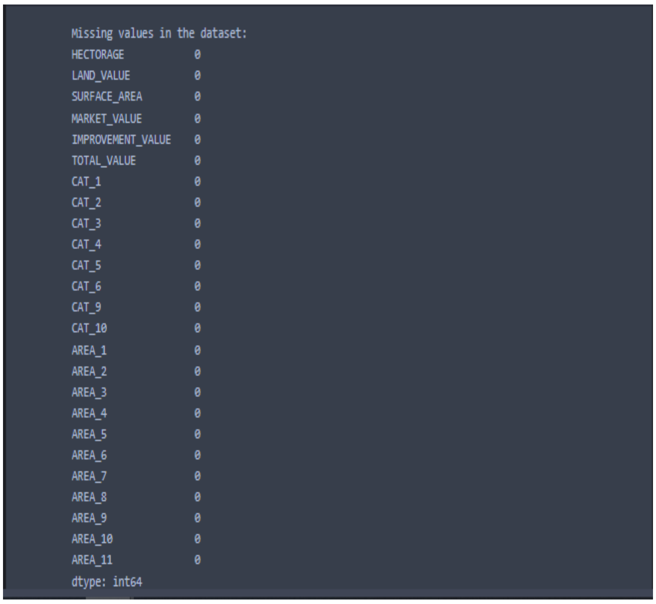


Fig.3. Missing values heatmap

The analysis of the dataset uncovered patterns of missing values among key variables, as illustrated by the heatmap visualization. Notably, SURFACE_AREA exhibited a moderate correlation of 0.2 with both HECTORAGE and LAND_VALUE, suggesting that missing values in SURFACE_AREA often coincided with missing values in these variables. Furthermore, MARKET_VALUE displayed a stronger correlation of 0.3 with HECTORAGE and a very strong correlation of 0.6 with IMPROVEMENT_VALUE, indicating significant overlap in their missing data. A detailed examination of the data confirmed these observations, revealing that SURFACE_AREA had 23 missing entries, MARKET_VALUE had 8, IMPROVEMENT_VALUE had 4, and LAND_VALUE had 1. These correlations indicated that the missingness was not completely random, but rather dependent on other variables, which informed the development of imputation strategies and impacted the robustness of subsequent analyses

To further investigate the nature of the missing data, Little's MCAR (Missing Completely At Random) test was performed. The results yielded a chi-square statistic of 384,286.24, accompanied by a p-value of 0.0, given 59,179 degrees of freedom. These findings strongly suggest that the missing data is not random, but rather dependent on other observed or unobserved factors. The rejection of the null hypothesis, which posits that data is missing completely at random, implies that the missingness in the dataset is likely influenced by other variables.

To address the issue of missing data, a K-Nearest Neighbors (KNN) imputation strategy was employed, followed by median or mode replacement for any residual missing data. This approach ensured that the dataset was complete and suitable for modeling. Additionally, categorical variables such as CAT and AREA underwent one-hot encoding, which enabled the machine learning algorithms to process the data effectively. The outcome of the imputation process is visualized in Fig. 4, demonstrating that the missing values were successfully handled



```
Missing values in the dataset:
HECTORAGE          0
LAND_VALUE         0
SURFACE_AREA       0
MARKET_VALUE       0
IMPROVEMENT_VALUE  0
TOTAL_VALUE        0
CAT_1              0
CAT_2              0
CAT_3              0
CAT_4              0
CAT_5              0
CAT_6              0
CAT_9              0
CAT_10             0
AREA_1             0
AREA_2             0
AREA_3             0
AREA_4             0
AREA_5             0
AREA_6             0
AREA_7             0
AREA_8             0
AREA_9             0
AREA_10            0
AREA_11            0
dtype: int64
```

Fig. 4 . Missing values after imputation

Outliers, identified and visualized using scatter plots as shown in Fig. 5, were mitigated through the application of log transformations. The effectiveness of this transformation is illustrated in Fig. 6, which demonstrates a significant improvement in the distribution of the data, resulting in a more normalized and symmetrical distribution. This transformation was particularly crucial for linear models, such as Linear Regression and Support Vector Regression (SVR), which assume a linear relationship between the predictor variables and the response variable. By reducing the impact of outliers, the log transformation enhanced the robustness and accuracy of these models, enabling a more reliable analysis of the relationships between the variables.

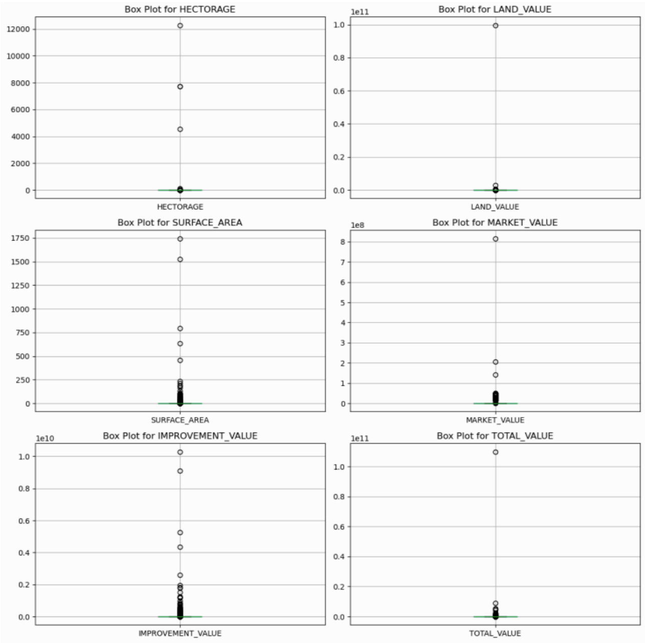


Fig. 5. Outliers

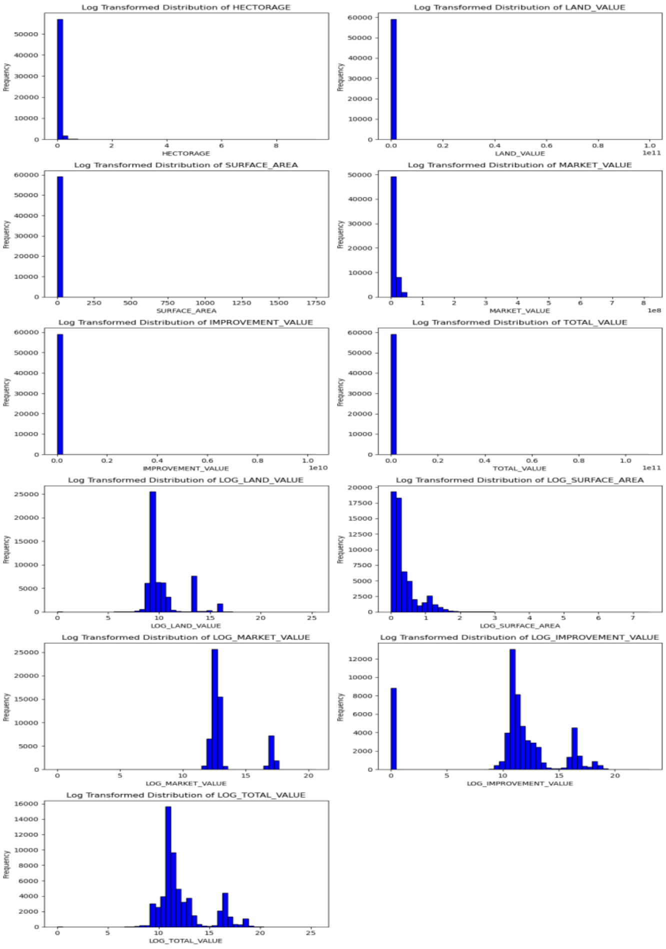


Fig. 6. Log Transformed data

D. *Model Develoment*

The Model Development adhered to the DRU Framework principles, ensuring theoretical rigor and practical applicability. The data was loaded in both transformed and raw forms to suit the requirements of different models, such as Linear Regression, SVM, Decision Trees, Random Forest, and Gradient Boosting. The and Evaluation Workflow included;

1. **Load Data**: Load the transformed dataset for linear regression and SVM, and the untransformed dataset for Decision Tree, Random Forest, and Gradient Boosting models.
2. **Train-Test Split**: Split data into training and testing sets to ensure unbiased evaluation, reserving the test set for final assessment.
3. **Validation Split**: Further split the training data into training and validation sets for parameter tuning and overfitting prevention.
4. **Feature-Target Separation**: Prepare independent variables (features) and dependent variable (target) for training, validation, and testing.
5. **Train Model**: Fit the model using training data (X_train, y_train).
6. **Make Predictions**: Generate predictions on validation (X_val) and test sets (X_test) to assess performance.

7. **Evaluate Performance**: Use validation data for tuning and test data for final evaluation, analyzing key metrics.
8. **Visualize Results**: Compare predicted vs. actual values and visualize performance metrics for interpretability.
9. **New Predictions**: Apply the trained model to unseen data for practical deployment.

E. *Model Evaluation*

The performance of the machine learning models was assessed based on three primary evaluation criteria: model accuracy, interpretability, and alignment with established real estate valuation principles. These criteria were selected to ensure that the models not only produced accurate predictions but also provided transparent and explainable results that conformed to the fundamental principles of real estate valuation.

1. Model Accuracy: was a crucial aspect of evaluating the performance of the machine learning models. To assess accuracy, we employed three key metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R²). The MSE and RMSE provided a measure of the average difference between predicted and actual values, with lower values indicating better fit.

a) Linear Regression Model fitting is illustrated in Fig.7 and Fig.8 The model metrics are out outlined in TABLE VI and TABLE VII.

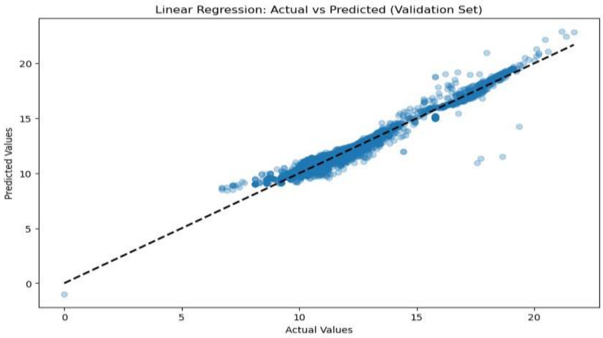


Fig.7. Linear Regression Validation

TABLE VI. LINEAR REGRESSION MODEL VALIDATION RESULTS

Validation Results			
Model	**MSE**	**RMSE**	**R²**
Linear Regression	0.1116	0.3342	0.9791

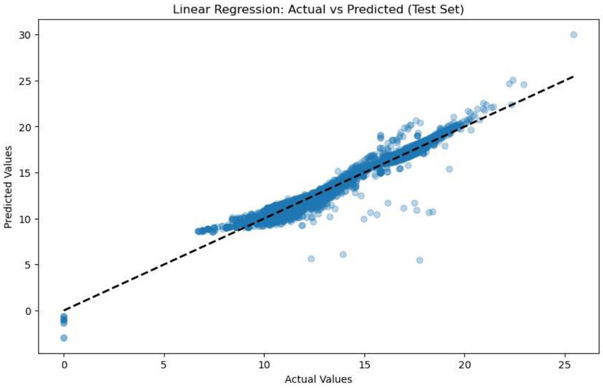


Fig. 8. Linear Regression Testing

TABLE VII. LINEAR REGRESSION MODEL TESTING RESULTS

Test Results			
Model	**MSE**	**RMSE**	**R²**
Linear Regression	0.1116	0.3342	0.9791

b) Decision Tree Model fitting is illustrated in Fig.9 and Fig.10 the model metrics are out outlined in TABLE III and TABLE IX.

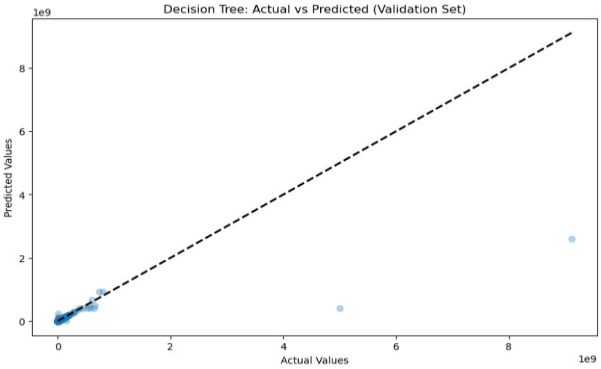


Fig.9. Decision Tree Validation

TABLE VIII. DECISION TREE MODEL VALIDATION RESULTS

Validation Results			
Model	**MSE**	**RMSE**	**R²**
Decision Tree	2.88E+13	5.37E+06	0.9736

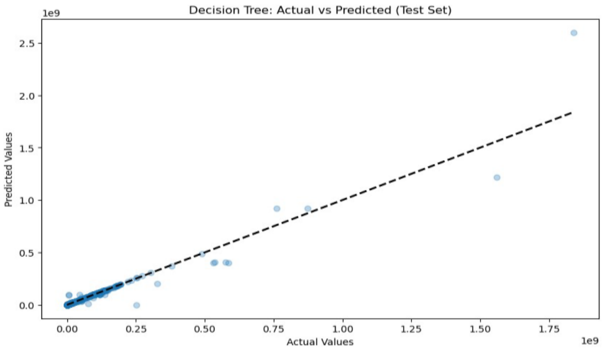


Fig. 10. Decision Tree Testing

TABLE IX. DECISION TREE MODEL TESTING RESULTS

Test Results			
Model	**MSE**	**RMSE**	**R^2**
Decision Tree	2.88E+13	5.37E+06	0.9736

c) Random Forest Model fitting is illustrated in Fig.11 and Fig.12 the model metrics are out outlined in TABLE X and TABLE XI.

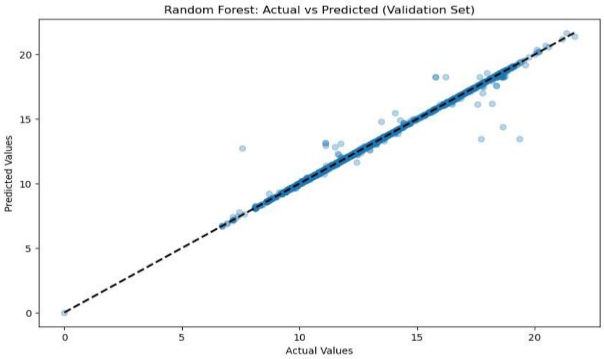


Fig.11. Randon Forest Validation

TABLE X. RANDOM FOREST MODEL VALIDATION RESULTS

Validation Results			
Random Forest	0.0101	0.1005	0.9981

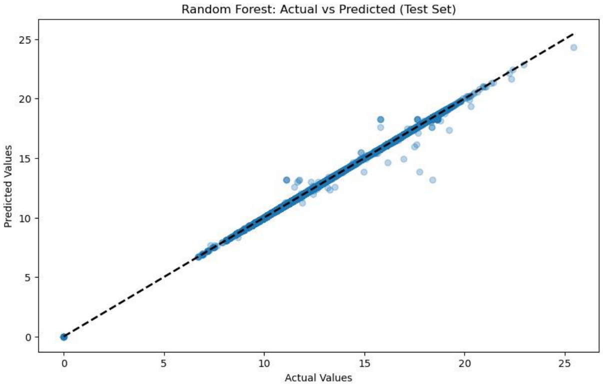


Fig.12. Randon Forest Testing

TABLE XI. RANDOM FOREST MODEL TESTING RESULTS

Test Results			
Model	**MSE**	**RMSE**	**R^2**
Random Forest	0.0026	0.051	0.9995

d) Gradient Boost Model fitting is illustrated in Fig.13 and Fig.14 the model metrics are out outlined in TABLE XII and TABLE XIII.

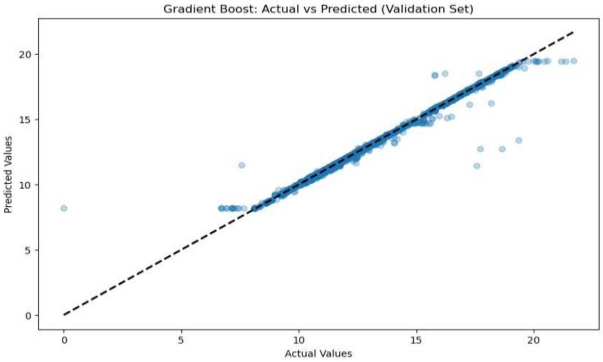


Fig.13. Gradient Boost Validation

TABLE XII. GRADIENT BOOSTING MODEL VALIDATION RESULTS

Validation Results			
Model	**MSE**	**RMSE**	**R^2**
Gradient Boosting	0.0204	0.1428	0.9961

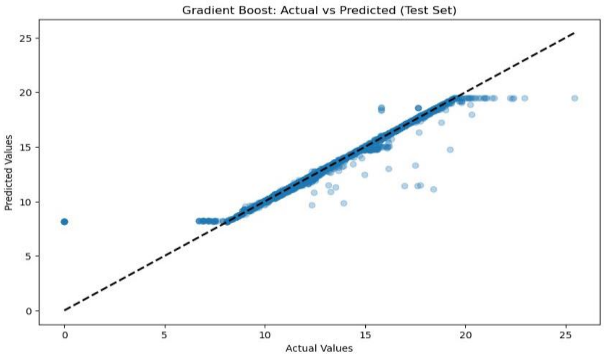


Fig.14. Gradient Boost Testing

TABLE XIII. GRADIENT BOOSTING MODEL TESTING RESULTS

Test Results			
Model	**MSE**	**RMSE**	**R²**
Gradient Boosting	0.0312	0.1767	0.9941

e) Support Vector Machine Model fitting is illustrated in Fig.15 and Fig.16 the model metrics are out outlined in TABLE XV and TABLE XIV.

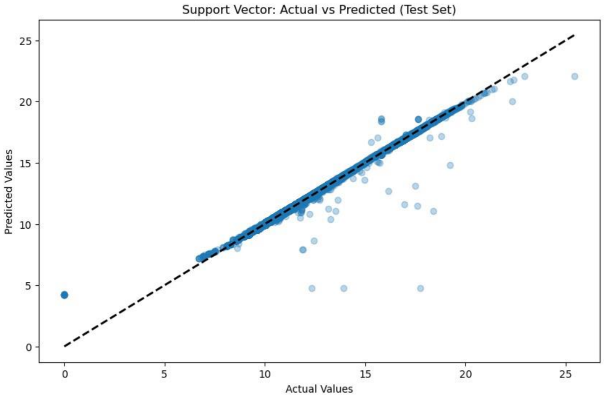


Fig.15. Support Vector Validation

TABLE XIV. SUPPORT VECTOR MODEL VALIDATION RESULTS

Validation Results			
Model	**MSE**	**RMSE**	**R²**
Support Vector Regression	0.0154	0.1242	0.997

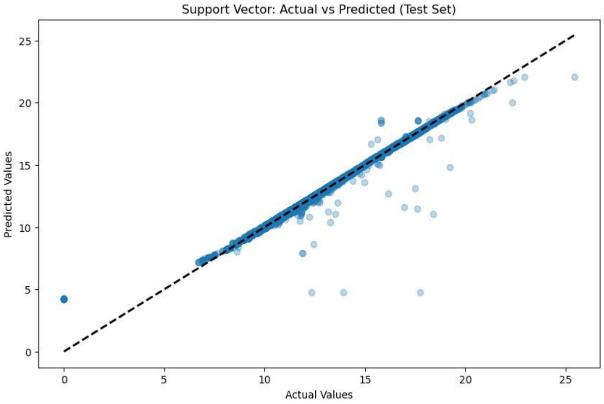


Fig.16. Support Vector Testing

TABLE XV. SUPPORT VECTOR MODEL TESTING RESULTS

Test Results			
Model	**MSE**	**RMSE**	**R²**
Support Vector Regression	0.022	0.1483	0.9958

The performance of each model was evaluated using three key metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²). These metrics provided a comprehensive view of each model's predictive accuracy and generalization capability.

Linear Regression consistently achieved an R² of 0.9791 on both the validation and test sets, indicating stable and reliable performance. However, its MSE of 0.1116 suggests limitations in capturing complex patterns, which may affect its suitability for datasets with non-linear relationships.

Decision Tree showed significant overfitting, as reflected by the sharp contrast between its validation R² of 0.4479 and test R² of 0.9736. The high MSE on the validation set (5.3975e+13) further underscores this model's poor generalization, making it a less reliable option for predictive tasks.

Random Forest emerged as the top performer, with R² values of 0.9981 on the validation set and 0.9995 on the test set. Its low test set MSE (0.0026) highlights its robustness and superior generalization capabilities, making it well-suited for complex datasets requiring high accuracy.

Gradient Boosting also demonstrated strong performance, with an R² of 0.9961 on the validation set and 0.9941 on the test set. Although slightly less accurate than Random Forest, its performance remains commendable, though the higher MSE suggests a slight risk of overfitting.

Support Vector Regression (SVR) achieved high accuracy with an R² of 0.9970 on the validation set and 0.9958 on the test set, and a test MSE of 0.0220. The model's ability to capture non-linear relationships makes it a robust choice for datasets with complex underlying patterns.

Fig. 17 provides a comparative analysis of MSE across all models, clearly illustrating the superior performance of the Random Forest model.

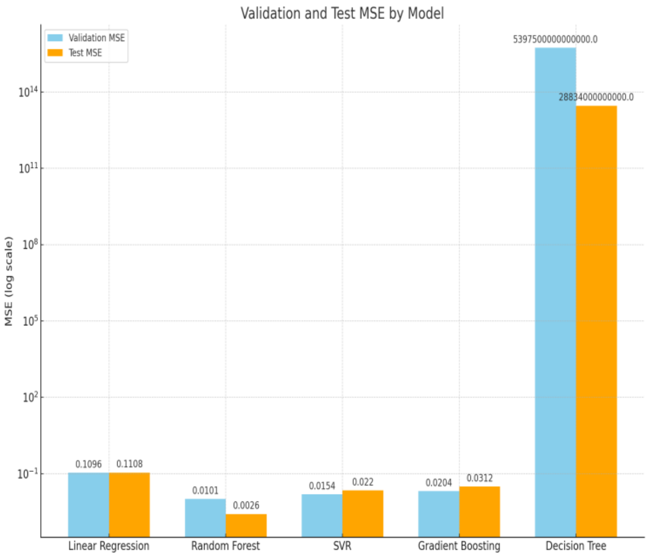


Fig.17. Comparing the all the models MSE

2. Explainability

The selection of machine learning models for property valuation required balancing predictive accuracy and explainability. Explainability, the extent to which a model's mechanics can be understood and trusted, was analyzed for five models: Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Regression (SVR).

Linear Regression was highly interpretable due to its assumption of a linear relationship between features and the target variable, allowing coefficients to directly quantify feature impact, though it struggled with non-linear relationships [32].

Decision Trees provided transparency by hierarchically dividing data, with feature importance assessed through splitting criteria, although deeper trees risked overfitting and reduced interpretability [33].

Random Forest improved accuracy by aggregating multiple Decision Trees and provided feature importance scores, but its ensemble nature made individual decision processes less transparent [34].

Gradient Boosting sequentially optimized errors to enhance accuracy, but its complexity and iterative nature limited full interpretability despite the availability of feature importance metrics [35].

SVR captured non-linear relationships using support vectors and kernel functions, offering limited insights due to the opacity introduced by kernel transformations [36].

Overall, Linear Regression and Decision Trees were the most explainable, making them suitable for cases prioritizing transparency. Random Forest and Gradient Boosting offered a balance between accuracy and explainability, while SVR, though effective for complex relationships, posed challenges for critical decision-making due to limited interpretability.

3. Consistency with real estate valuation principles

Supply and Demand Theory**:** The relationship between supply, demand, and property value was captured by features like *LAND_VALUE*, *MARKET_VALUE*, and *IMPROVEMENT_VALUE*. These reflected market conditions influencing property values. For example, *MARKET_VALUE* was impacted by external economic factors, aligning with Marshall’s supply and demand theory. Models such as Linear Regression and Random Forest used these variables to predict property values based on market forces [1].

Highest and Best Use Theory: The Highest and Best Use principle was represented by *HECTORAGE*, *SURFACE_AREA*, and *CATEGORY*. These features helped determine the optimal use of a property, with Decision Trees and Random Forest models identifying key attributes influencing property value based on intended use and market demand. *HECTORAGE* and *CATEGORY* distinguished between land uses, while *SURFACE_AREA* aided in assessing development potential [2].

Marginal Utility Theory: Marginal utility, which suggested diminishing returns with additional features, was reflected in *IMPROVEMENT_VALUE* and *LAND_VALUE*. These variables modeled the diminishing contribution of property features like land size or development as they increased. Gradient Boosting and SVR captured this non-linear relationship [3].

Comparable Market Analysis: *MARKET_VALUE* and *AREA* represented the property’s value relative to similar properties and geographic location. Random Forest and Gradient Boosting models analyzed market trends, location, and property category to replicate Comparable Market Analysis, using these features to estimate property values based on comparable data [4].

Property-Specific Characteristics: Variables such as *HECTORAGE*, *SURFACE_AREA*, *LAND_VALUE*, and *IMPROVEMENT_VALUE* were critical in assessing a property’s unique value. Linear Regression, Decision Trees, and Random Forest incorporated these features to evaluate tangible and intangible factors that influenced property worth [5].

Valuation Standards and Approaches: The models aligned with established valuation standards from the Appraisal Institute. Metrics like *LAND_VALUE*, *MARKET_VALUE*, *IMPROVEMENT_VALUE*, and *HECTORAGE* were integral to methods like the sales comparison and income capitalization approaches, ensuring consistency with traditional valuation practices while leveraging modern machine learning techniques [4].

V. DISCUSSION

The results of this study underscore the significant potential of machine learning models in transforming property valuation, particularly in dynamic and complex real estate markets where traditional methods often fall short. Both the Random Forest and Support Vector Regression (SVR) models demonstrated exceptional performance, yielding notably low Mean Squared Error (MSE) and high R-squared (R^2) values. These results validate the models' effectiveness in managing intricate datasets that incorporate diverse property features such as LAND_VALUE, MARKET_VALUE, and IMPROVEMENT_VALUE. The high accuracy achieved suggests that municipalities can confidently adopt these models for data-driven decisions, enhancing financial stability through precise property tax assessments.

This outcome aligns with the primary objectives of the study, which are to assess the accuracy and interpretability of machine learning models in property valuation, ensuring that predictions are both accurate and explainable. Moreover, the study aims to verify that model predictions align with established real estate principles and to identify key variables in property valuation, including LAND_VALUE, MARKET_VALUE, and IMPROVEMENT_VALUE.

The Random Forest model excels due to its capacity to capture non-linear relationships and interactions between variables, while the SVR model's strength lies in its flexibility and robustness in mapping input-output relationships. These attributes are particularly valuable in real estate markets, which are influenced by economic trends, location-specific factors, and government policies. The findings, therefore, not only highlight the models' predictive power but also emphasize their alignment with real estate principles.

Additionally, the adoption of the Data, Reasoning, and Usefulness (DRU) Evaluation Framework adds rigor to this research. By integrating the Technology Acceptance Model (TAM) and Task Technology Fit (TTF) model, the study evaluates both the predictive performance of the models and their practical alignment with municipal governance needs. This comprehensive framework ensures that the models are not only accurate but also interpretable and actionable, bridging the gap between technical innovation and real-world application.

Despite the promising results, some limitations persist. The accuracy of the models is contingent upon the quality of the input data, and the computational demands of Random Forest and SVR models can be considerable. However, these challenges can be mitigated through cloud-based solutions or hybrid approaches that balance computational efficiency and model accuracy. Additionally, geographic variability must be carefully considered, as a model trained in one region may not generalize well to another due to differences in market dynamics, land-use patterns, and property characteristics.

In conclusion, the study's findings demonstrate the transformative potential of machine learning models for property valuation. While acknowledging the challenges associated with these models, the study emphasizes their alignment with real estate principles, offering municipalities a reliable tool for making data-driven decisions that improve financial stability and property valuation accuracy.

VI. Conclusion

Accurate property valuation is a cornerstone of municipal governance, directly impacting tax assessments and urban planning. This study highlights the potential of machine learning models, particularly Random Forest and SVR, to overcome the limitations of traditional valuation methods. By achieving exceptional predictive performance, these models offer municipalities a robust framework for making fair and informed decisions.

The research not only demonstrates the technical superiority of machine learning but also emphasizes its practical implications. The application of advanced regression techniques can streamline valuation processes, reduce manual interventions, and adapt to market dynamics, thus fostering financial stability and equity in tax assessments. Additionally, the integration of the DRU Evaluation Framework ensures that the models provide actionable insights while maintaining alignment with real estate valuation principles.

Despite their promise, the performance of these models is contingent upon high-quality data inputs and adequate computational resources. Future research should focus on addressing these challenges by exploring hybrid models, incorporating real-time market data, and validating model generalizability across diverse regions. Such advancements would further solidify the role of machine learning in revolutionizing property valuation practices.

By investing in continuous data updates and leveraging advanced computational tools, municipalities can harness the full potential of machine learning to enhance governance and urban planning. This research serves as a critical step toward modernizing property valuation, paving the way for more accurate, equitable, and efficient municipal processes.

References

[1] Ding, C., & Hwang, J. (2018). Machine learning in property valuation: An application of random forest regression. Journal of Property Research, 35(4), 381-397.

[2] Gao, J., &Asami, Y. (2011). Preferences for land readjustment projects in urban China: A comparison between residents and experts. Urban Studies, 48(5), 1123-1140.

[3] Kok, N., Koponen, E.-L., &Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. Journal of Property Investment & Finance, 35(2), 135-153.

[4] González, V. M., &Formoso, C. T. (2006). Mass appraisal with genetic fuzzy systems: An application for the residential real estate market. Expert Systems with Applications, 30(2), 296-306

[5] T. H. Root, T. J. Strader, and Y.-H. (John) Huang, "A review of machine learning approaches for real estate valuation," *Journal of the Midwest Association for Information Systems*, vol. 2023, no. 2, Art. no. 2, 2023, doi: 10.17705/3jmwa.000082.

[6] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.

[7] D. Goodhue and R. L. Thompson, "Task-technology fit and individual performance," *MIS Quarterly*, vol. 19, no. 2, pp. 213–236, 1995.

[8] D. Gefen and D. W. Straub, "The relative importance of perceived ease of use in IS adoption: A study of e-commerce adoption," *Journal of the Association for Information Systems*, vol. 1, no. 1, Art. 8, 2000.

[9] S. Rosen, "Hedonic prices and implicit markets: Product differentiation in pure competition," Journal of Political Economy, vol. 82, no. 1, pp. 34-55, 1974. 1

[10] N. Nguyen and A. Cripps, "Predicting housing value: A comparison of multiple regression analysis and artificial neural networks," Journal of Real Estate Research, vol. 22, no. 3, pp. 313-336, 2001.

[11] Din, M. Hoesli, and A. Bender, "Environmental variables and real estate prices," Urban Studies, vol. 38, no. 11, pp. 1989-2000, 2001.

[12] J. H. Chen, T. T. Chang, C. R. Ho, and J. F. Diaz, "Grey relational analysis and neural network forecasting of REIT returns," Quantitative Finance, vol. 14, no. 11, pp. 2033-2044, 2014.

[13] E. Pagourtzi, V. Assimakopoulos, T. Hatzichristos, and N. French, "Real estate appraisal: A review of valuation methods," Journal of Property Investment & Finance, vol. 21, no. 4, pp. 383-401, 2003.

[14] M. McCluskey, M. McCord, P. T. Davis, M. Haran, and D. McIlhatton, "Prediction accuracy in mass appraisal: A comparison of modern approaches," Journal of Property Research, vol. 30, no. 4, pp. 239-265, 2013.

[15] V. Kontrimas and A. Verikas, "The mass appraisal of the real estate by computational intelligence," Applied Soft Computing, vol. 11, no. 1, pp. 443-448, 2011.

[16] M. Cajias and S. Ertl, "Spatial effects and non-linearity in hedonic modeling: Will large data sets change our assumptions?" Journal of Property Investment & Finance, vol. 36, no. 1, pp. 32-49, 2018.

[17] S. Mullainathan and J. Spiess, "Machine learning: An applied econometric approach," Journal of Economic Perspectives, vol. 31, no. 2, pp. 87-106, 2017.

[18] J. A. Yacim and D. G. B. Boshoff, "Impact of artificial neural networks training algorithms on accurate prediction of property values," Journal of Real Estate Research, vol. 40, no. 3, pp. 375-418, 2018.

[19] J. A. Yacim and D. G. B. Boshoff, "Combining BP with PSO algorithms in weights optimisation and ANNs training for mass appraisal of properties," International Journal of Housing Markets and Analysis, vol. 11, no. 2, pp. 290-314, 2018.

[20] J. Wen, S. Li, Z. Lin, Y. Hu, and C. Huang, "Systematic literature review of machine learning based software development effort estimation models," Information and Software Technology, vol. 54, no. 1, pp. 41-59, 2012.

[21] N. Marangunić and A. Granić, "Technology acceptance model: A literature review from 1986 to 2013," Universal Access in Information Society, vol. 14, pp. 81-95, 2015.

[22] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data quality in context," Communications of the ACM, vol. 40, no. 5, pp. 103-110, 1997.

[23] B. Furneaux, "Task-technology fit theory: A survey and synopsis of the literature," Information Systems Theory: Explaining and Predicting Our Digital Society, vol. 1, pp. 87-106, 2012.

[24] D. L. Goodhue and R. L. Thompson, "Task-technology fit and individual performance," MIS Quarterly, vol. 19, no. 2, pp. 213-236, 1995.

[25] W. Ho, B. Tang, and S. Wong, "Predicting property prices with machine learning algorithms," Journal of Property Research, vol. 38, no. 1, pp. 48-70, 2021.

[26] J. Hong, H. Choi, and W. Kim, "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, vol. 24, no. 3, pp. 140-152, 2020.

[27] R. Spies, S. Grobbelaar, and A. Botha, "A scoping review of the application of the task-technology fit theory," in Responsible Design, Implementation and Use of Information and Communication Technology, pp. 397-408, 2020.

[28] Bogin and J. Shui, "Appraisal accuracy and automated valuation models in rural areas," Journal of Real Estate Finance and Economics, vol. 60, pp. 40-52, 2020.

[29] J. Rico-Juan and P. Taltavull de La Paz, "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, vol. 171, pp. 1-14, 2021.

[30] J. J. Ahn, H. W. Byun, K. J. Oh, and T. Y. Kim, "Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting," Expert Systems with Applications, vol. 39, no. 9, pp. 8369-8379, 2012.

[31] M. Cajias and S. Ertl, "Spatial effects and non-linearity in hedonic modeling: Will large data sets change our assumptions?" Journal of Property Investment & Finance, vol. 36, no. 1, pp. 32-49, 2018.

[32] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed. New York, NY, USA: Wiley-Interscience, 1998.

[33] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* Belmont, CA, USA: Wadsworth, 1986.

[34] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[35] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[36] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York, NY, USA: Springer, 1999.