Available online at
https://mijuitm.com.my

e-ISSN: 2735-0703

Mathematical Sciences and Informatics Journal 6(2) 2025, 286 – 304.

Mathematical Sciences and Informatics Journal

Predicting Real Estate Prices with AI: A Comparative Study of Machine Learning Models

Eng-Lian Lim[1*], Doris Hooi-Ten Wong[1] and Maslin Masrom[1]

[1]*Universiti Teknologi Malaysia*

ARTICLE INFO

Article history:
Received 10 August 2025
Revised 23 August 2025
Accepted 19 September 2025
Online first
Published 31 October 2025

Keywords:
Machine Learning
House Price Prediction
Data-Driven Approach
Random Forest
Gradient Boosting Machine,
Artificial Neural Networks
Real Estate,
CRISP-DM

DOI:
10.24191/mij.v6i2.9178

ABSTRACT

Accurate house price prediction is vital for economic, financial, and policy decision-making, impacting homebuyers, investors, financial institutions, and government agencies. This study employs data-driven machine learning approach to forecast residential property prices, with a particular focus on high-rise properties in Kuala Lumpur. Real-world housing data comprising 12,735 transactions (2021–August 2024) were collected from the National Property Information Centre (NAPIC), preprocessed, and analyzed using exploratory data analysis (EDA) to understand the influence of various property attributes on prices. Multiple predictive models, including traditional regression, ensemble methods (Random Forest, Gradient Boosting Machines), and deep learning (Artificial Neural Networks), were developed and rigorously compared. Model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 on an 80:20 training-testing split. Hyperparameter tuning and K-fold cross-validation were applied to optimize accuracy, prevent overfitting, and ensure model generalizability. The Random Forest model was the best-performing predictor, with the lowest error values and highest R^2 score compared to other tested algorithms. This research offers practical insights into key price-influencing features and highlights the efficacy of machine learning for robust and interpretable house price prediction in the Malaysian real estate market.

1. INTRODUCTION

Accurate forecasting of residential property values is vital for stakeholders, including homebuyers, developers, financial institutions, and policymakers, who rely on reliable predictions for investment, risk assessment, lending, and urban planning. However, traditional market analyses, often based on broad economic indicators, struggle to capture the non-linear interplay of factors like location, architectural features, market sentiment, and regulatory shifts that driving housing prices, particularly in dynamic urban markets like Kuala Lumpur's high-rise sector (Ravikumar & Nagashree, 2023). Fig. 1 illustrates Malaysia's property price index from 2010 to 2024, showing a sustained rise in average house prices, more than

[1*] Corresponding author. *E-mail address*: limenglian@graduate.utm.my
https://doi.org/10.24191/mij.v6i2.9178

doubling over the period, though annual growth slowed significantly after 2012, reaching low single-digit rates by 2019–2021, with a modest recovery recently and a projected slight moderation in 2024.

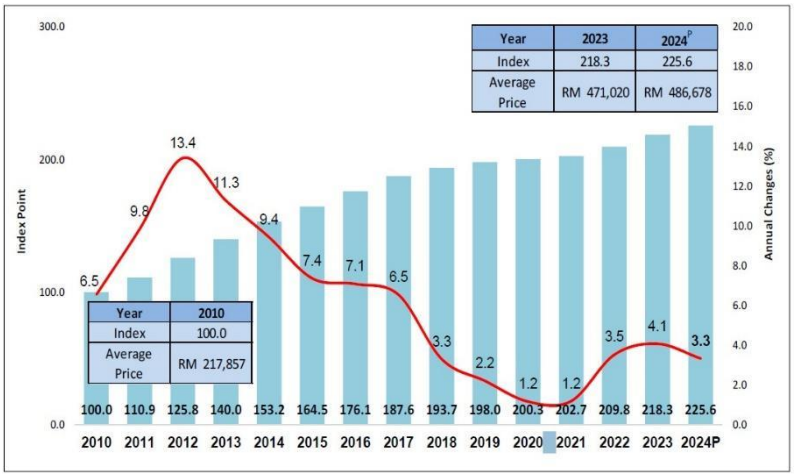

Fig. 1. MHPI index point and annual change 2010 – 2024P

Prior studies using public datasets have improved predictive modeling, but limited data scope often reduces real-world applicability (Tahir & Fatima, 2024). In Malaysia, research such as (Ja'afar et al., 2021) and (Sa'at et al., 2021) applied machine learning to house price prediction but typically used smaller datasets or relied on traditional methods like hedonic pricing, which struggle with complex market dynamics. This study introduces a novel data-driven approach by leveraging a comprehensive dataset of 12,735 high-rise residential transactions from the National Property Information Centre (NAPIC, 2021-2024), enabling robust modeling of Kuala Lumpur's diverse high-rise market. Unlike earlier studies, this research systematically compares a broad range of machine learning models: Multiple Linear Regression, Decision Trees, Random Forest, Gradient Boosting Machine, Support Vector Machine, and Artificial Neural Networks within the CRISP-DM framework. Through advanced feature engineering, including logarithmic transformations and categorical encoding, and rigorous hyperparameter tuning to enhance accuracy and interpretability, this research advances prior work to fill a critical gap by providing a scalable and interpretable predictive framework tailored to Kuala Lumpur's high-rise market, overcoming limitations of prior studies with restricted data or simpler models. This study proposes a data-driven approach to predict house prices under stable macroeconomic conditions, using a comprehensive dataset to reflect market behavior (Gulati & Raheja, 2021). It emphasizes data quality through rigorous preprocessing, including cleansing, transformation, normalization, and outlier management (Ved & Gupta, 2024). While neural networks show promise for price forecasting, their potential in local markets remains underexplored due to reliance on simplified datasets. The paper is structured as follows: Section 2 reviews literature on machine learning, feature selection, and model interpretability. Section 3 details the methodology, Section 4 presents results and findings, and Section 5 offers conclusions and future recommendations.

2. LITERATURE REVIEW

The field of real estate valuation has evolved significantly, with traditional methods forming the foundation and machine learning approaches addressing their limitations. To enhance clarity, this section is organized into three parts: traditional valuation methods, machine learning approaches, and identified research gaps.

https://doi.org/10.24191/mij.v6i2.9178

2.1 Traditional Valuation Methods

The field of real estate valuation has historically relied on a variety of methods to assess property values (StarProperty, 2024). Understanding these established approaches is crucial for appreciating the evolution towards more sophisticated, data-driven techniques.

Traditional valuation methods often serve as foundational tools, but they also come with inherent limitations. Comparative market analysis (CMA) involves comparing a subject property to recently sold similar properties in the same area, adjusting for differences. While intuitive and widely used, its accuracy heavily depends on the availability of truly comparable sales and the expertise of the appraiser in making subjective adjustments.

$$Estimated\ Property\ Value\ = \frac{\left(\Sigma\ (Sale\ Price\ \pm\ Adjustments)\right)}{n} \tag{1}$$

where:

$Sale\ Price$: The actual price at which a similar property was sold.
$Adjustments$: Dollar value changes to account for differences in size, location, age, number of bedrooms, renovations, etc.
n : Number of comparable properties used.

The cost approach estimates property value by summing up the cost of the land and the depreciated cost of replacing or reproducing the building. This method is particularly useful for new constructions or unique properties where comparable sales are scarce, but it can be challenging to accurately estimate depreciation and construction costs in rapidly changing markets (Chan et al., 2023).

$$Property\ Value\ =\ Land\ Value\ +\ (Cost\ to\ Rebuild\ New\ -\ Depreciation) \tag{2}$$

The income approach is primarily used for income-generating properties, capitalizing the property's net operating income into a value estimate. Its effectiveness is tied to stable income streams and accurate market capitalization rates, making it less suitable for residential properties or volatile markets (Rattanaprichavej & Teeramungcalanon, 2020). Lastly, the Expert Appraisal Method relies on the professional judgment of a certified appraiser, integrating various market factors and personal experience. While offering a holistic view, its subjectivity can lead to inconsistencies across different appraisals. These traditional methods, while valuable, often struggle to capture complex, non-linear relationships and numerous influencing factors that characterize modern real estate markets.

2.2 Machine Learning Approaches

In recent decades, there has been a significant shift towards data-driven approaches, particularly machine learning, to overcome the limitations of conventional valuation techniques (Chen, 2024). Machine learning offers the ability to process vast amounts of data and identify intricate patterns and relationships that are not readily apparent through linear models or human intuition. Common machine learning algorithms applied in real estate price prediction include traditional regression models such as Linear Regression, which establish a linear relationship between features and price; Decision Trees, which make predictions based on a series of decision rules derived from data features; and more advanced ensemble methods. Ensemble learning, such as Random Forest and Gradient Boosting Machines (GBM), combines multiple individual models to produce a more robust and accurate prediction, excelling at capturing complex interactions and non-linearities in data (McCluskey et al., 2014). Furthermore, deep learning models like Artificial Neural Networks (ANN) have shown considerable promise by learning hierarchical features from raw data, potentially uncovering deeper, more abstract patterns in real estate data. A critical aspect of applying machine learning in real estate is effective feature selection and engineering. This involves identifying the most influential property attributes (e.g., location, size, number of rooms,

https://doi.org/10.24191/mij.v6i2.9178

amenities, age) and transforming raw data into features that enhance model performance. For instance, creating new features from existing ones (e.g., age of property from construction year) or handling categorical variables through encoding are crucial steps (Rahmat et al., 2023).

2.3 Research Gaps

Despite the growing adoption of machine learning, a comparative analysis of prior studies reveals persistent limitations. Many early studies relied heavily on time series forecasting models, such as ARIMA, which are often limited in their ability to integrate diverse housing attributes or perform robustly in volatile and dynamic market conditions. Such models typically focus on historical price trends rather than the multifaceted factors influencing individual property values. Additionally, research exploring the impact of emerging features, such as green certification or properties within gated communities, frequently utilize small, geographically confined datasets. This narrow scope inherently limits the scalability and generalizability of their findings to broader or more diverse real estate markets. Even with the demonstrated efficacy of ensemble tree-based methods like Boosted Regression Trees (McCluskey et al., 2014), many studies, particularly those focusing on the Malaysian housing market, have continued to employ relatively small and geographically restricted datasets.

A systematic review of machine learning applications in property price prediction (Ja'afar et al., 2021) highlighted that while Random Forest was a prevalent method between 2011 and 2019, a significant drawback was the lack of spatial diversity in most implementations, consequently hindering their generalizability across different regions. These identified limitations collectively underscore a pressing need for more robust, interpretable, and broadly generalizable modeling approaches in real estate valuation. This review leads to the establishment of the context for the systematic and comprehensive data-driven methodology employed in this research, which adheres to the CRISP-DM framework for data preparation, preprocessing, feature selection & engineering, model training, evaluation, and deployment. Table 1 lists a few previous studies on house price prediction for the Malaysian housing market from 2021 until 2025.

Table 1 Previous studies on house price prediction for the Malaysian housing market (2021 – 2025)

No	Year	Title	Model	Significance of finding	Limitation
1	2025	Empirical forecasting of housing prices in Malaysia using ARIMA.	ARIMA	MHPI predictions for different house types from 2025 - 2030 Fitted models for all houses: ARIMA (0,1,0)	Linear relationship assumed Macroeconomic factors and other spatial attributes were excluded from the analysis
2	2024	Malaysian Residential Property: A Forecasting Model	ARIMA	Interest rates have one way effect on MHPI ARIMA (1,2,1) is the best model for MHPI prediction.	One macroeconomic factor studied for its importance to HPI prediction, but cannot be included in the model
3	2024	The Effect of Security in the Green Building Price Prediction Model	Multiple Linear Regression Vs. Random Forest, Decision Tree, Linear Regressor, Ridge, and Lasso Regressor	Correlation Analysis shows security is a significant factor & MLR indicated that security has a negative correlation to house price, unexpectedly. ML methods failed to validate the MLR outcomes	The data set used for the study is relatively small: 240 records of green buildings in Kuala Lumpur.
4	2023	Supervised Machine Learning Approach to Housing Market	Linear Regression, KNN (for attribute classification	The high MSE and MAE values indicate substantial prediction errors, while the R-squared value suggests that the model could benefit from further refinement or exploration of more complex modelling approaches	Linear relationship assumed
5	2023	Forecasting the Volatility of Real Residential	Generalised Autoregressive Conditional	Time series model explaining house price volatility. The result reveals stability in house prices at the beginning of 2023.	Not including important spatial attributes and macroeconomic factors in the

https://doi.org/10.24191/mij.v6i2.9178

		Property Prices in Malaysia	Heteroskedasticity (GARCH) model		prediction model. House prices usually don't have rapid volatility spikes
6	2022	Heritage Properties Price Prediction Using Random Forest Classifier	Random Forest	Random Forest model produced better performance on the heritage property	Scope at heritage properties only.
7	2022	Hedonic Regression Analysis in determining the effect of Green on high-rise residential	Hedonic Regression Analysis (Multiple Linear Regression)	A Green certificate has an impact on condominium prices in the Timur Laut district, Penang. A GBI-certified house has a 6.5% premium in price over a Non-GBI-Certified house.	Linear relationship assumed Model accuracy is low at 0.36
8	2021	Enhancing the Accuracy of Malaysian House Price Forecasting	Hedonic Price Model (HPM) vs Artificial Neural Network (ANN)	ANN outperformed HPM with lower RMSE	A linear relationship is assumed for the HPM model Only a few attributes are included. % of error is more than 10
9	2021	Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur	LightGBM, XGBoost Vs. Multiple Regression Analysis (MRA), Ridge Regression	House price analysis by region is chosen due to price dependence on location. Locational variables included: distance to shopping mall, school, hospital, LRT/MRT station XGBoost outperformed other models	Limited to the KL region, i.e., broader locational attributes are excluded.
10	2021	Machine Learning for Property Price Prediction and Price Valuation: A Systematic Literature Review	Random Forest, Decision Tree, Gradient Boosting, Neural Network and Linear Regression	Random Forest is the most selected method for house price prediction among other models from 2011-2019	No empirical results were provided in the study

3. METHODOLOGY

This study adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework (Martínez-Plumed et al., 2021), a widely recognized iterative methodology for data science projects, to ensure a systematic and replicable approach. The framework, as illustrated in Fig. 2 below, organizes this project into six phases:

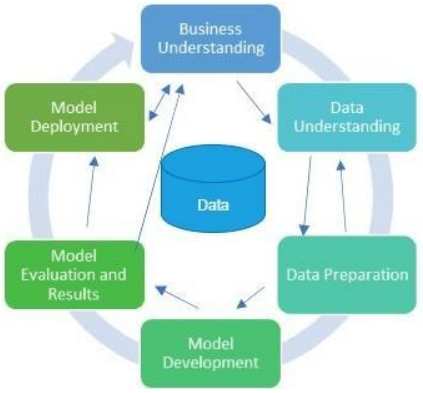

Fig. 2. CRISP-DM framework and processes

https://doi.org/10.24191/mij.v6i2.9178

3.1 Business Understanding

This initial phase focused on clearly defining the research objectives from a business perspective. The primary goal was to develop a robust and accurate model for predicting high-rise residential property prices in Kuala Lumpur. This involved understanding the specific needs of stakeholders such as homebuyers, investors, financial institutions, and government agencies, who require reliable price forecasts for informed decision-making. The project aimed to address the challenges posed by market complexities, inflation, and changing buyer behaviors by providing a data-driven solution that surpasses traditional valuation methods in accuracy and interpretability.

3.2 Data Understanding

Real-world high-rise residential property transaction data for Kuala Lumpur from 2018 to 2023 was acquired from the National Property Information Centre (NAPIC). The dataset comprises 428,799 transactions and includes variables such as property type, transaction date, price, floor area, tenure, and location (Mukim). Exploratory Data Analysis (EDA) was conducted to gain insights into data distributions, identify potential outliers, and understand preliminary relationships between variables and transaction prices. Time series analysis reveals generally increasing trends in property prices over the study period, with some seasonal variations.

3.3 Data Preparation

This phase involved comprehensive data cleaning and transformation. Steps included handling missing values through imputation or removal, identifying and treating outliers, and converting categorical features into numerical representations using label encoding. Feature engineering was also performed to create new, more informative features from existing ones, for example, property age from the transaction date. Data normalization and standardization techniques were applied to ensure consistent scales across features, crucial for model performance.

3.4 Model Development

The prepared dataset was split into an 80% training set and a 20% testing set. Various machine learning models were developed and trained, including Multiple Linear Regression (MLR), Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machine (SVM) and Artificial Neural Networks (ANN). Each model was selected for its specific strengths in addressing the complexities of house price prediction, as outlined below:

MLR is a fundamental supervised learning algorithm that assumes the price of a house can be linearly explained by multiple independent variables (features) such as land/parcel area, location, and year of tenure. It relies on assumptions like linearity, independence of errors, homoscedasticity, normal distribution of error terms, and absence of multicollinearity among predictors. Its primary advantage is interpretability, allowing quantification of the individual effect of each independent variable on the dependent variable. It is also computationally efficient and suitable for small to moderate-sized datasets (Bau & Hisham, 2022). MLR is sensitive to outliers, struggles to capture complex nonlinear relationships, and its performance may decrease in high-dimensional datasets with correlated predictors. MLR was chosen for its simplicity and interpretability, serving as a baseline for comparison.

Decision Trees are among the simplest machine learning models that split data into branches based on feature values to maximize homogeneity of the target variable (house price in this study) within each branch. The prediction at a leaf node is typically the average of target values for all training samples in that node. Decision trees provide a measure of feature importance, reflecting the magnitude of MSE reduction at each split. Decision Trees were selected for their ability to handle non-linear relationships and provide feature importance insights. Random Forests ensemble learning techniques that build multiple decision

https://doi.org/10.24191/mij.v6i2.9178

trees, each trained on a random subset of the data. They aggregate the predictions of these multiple trees to enhance accuracy and reduce variance. RF is one of the most widely used and effective methods for house price prediction, demonstrating robustness in handling non-linearity and high-dimensional data. Random Forest was chosen for its robustness and high accuracy in capturing complex interactions in real estate data.

GBMs are powerful ensemble learning techniques that build models sequentially by optimizing a loss function through Gradient Descent (Chen & Guestrin, 2016). Each new tree attempts to correct the residual errors made by the previous one. GBM models are effective in capturing complex, nonlinear relationships and have shown superior performance with structured and tabular datasets (e.g., XGBoost, LightGBM). Many recent studies (Ja'afar & Mohamad, 2021; Chun et al., 2025) have demonstrated the outstanding performance of tree-based models and their ensembled variants in house price prediction. GBM was selected for its ability to model non-linear patterns and improve predictive performance through sequential learning. SVMs are supervised learning models for classification and regression. SVR is a variation used for regression tasks like house price prediction, aiming to fit data within a specified margin of tolerance (epsilon) while minimizing error and maintaining generalization. SVR demonstrated solid predictive performance in real estate markets as evidenced by Wang et al. (2021) and Ho et al. (2020). SVM was chosen for its effectiveness in handling high-dimensional data and robustness to outliers.

ANNs are computational models inspired by the human brain, consisting of interconnected nodes or "neurons" arranged in layers that process information through weighted connections. Typically comprises an Input Layer, one or more Hidden Layers (where computations and feature transformations occur using activation functions), and an Output Layer producing the final prediction. It is particularly suited for modeling non-linear and complex relationships in large datasets, leading to rapid adoption in house price predictions (Sa'at et al., 2021). However, it is often characterized as a "black box" due to its inherent complexity, making predictions challenging to interpret or explain to stakeholders. ANN was chosen to explore its potential in capturing intricate, non-linear patterns in large-scale real estate data.

3.5 Model Evaluation

Model performance was rigorously evaluated using standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) as shown in Equations 3 until 6. To enhance model generalization and prevent overfitting, K-fold cross-validation was employed. Hyperparameter tuning, particularly for ensemble and neural network models, was conducted using techniques like GridSearchCV to identify optimal parameter settings that maximize predictive accuracy.

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

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}\left(y_i - \underline{y}\right)^2} \tag{6}$$

https://doi.org/10.24191/mij.v6i2.9178

3.6 Model Deployment

The best-performing model, once validated, will be saved using appropriate serialization techniques (e.g., Python's pickle library) for potential future deployment in real-world applications or integration into a web-based prediction tool.

3.7 Pipeline Summary, Tools, and Platform

The core workflow involves environment setup, data preparation (loading, preprocessing, cleaning, transformation), feature analysis and engineering (encoding, outlier handling, log transformations), model training (various regressors), model evaluation (RMSE, R-squared), model tuning (hyperparameter optimization), and finally, model deployment (saving and implementing prediction functions).

Fig. 3 illustrates the machine learning pipeline for this regression task. The process systematically progresses from environment setup and library import, through data loading and preprocessing, to feature engineering for a training-ready dataset. Subsequently, various regressors are trained, evaluated using performance metrics, and optimized through hyperparameter tuning. Finally, the best models are deployed for future use and predictions, ensuring an efficient and repeatable workflow. The research was primarily conducted using Python programming language. Key libraries included Pandas and NumPy for data manipulation, Scikit-learn for machine learning algorithms, Matplotlib and Seaborn for data visualization, and TensorFlow/Keras for deep learning models. Google Colab and Visual Studio Code served as the primary development environments.

Table 2 lists the tools, platforms and resources that were used in the study. Firstly, the research was conducted using Python (version 3.12.11) as the primary programming language, with key libraries including Pandas (2.2.2) and NumPy (2.0.2) for data manipulation, Scikit-learn (1.2.2) for machine learning algorithms, then, Matplotlib (3.10.0) and Seaborn (0.13.2) were implemented for data visualization, and TensorFlow/Keras (2.19.2) for deep learning models. Finally, Google Colab 0.0.1a2 and Visual Studio Code (version 1.85) served as the primary development environments to ensure reproducibility.

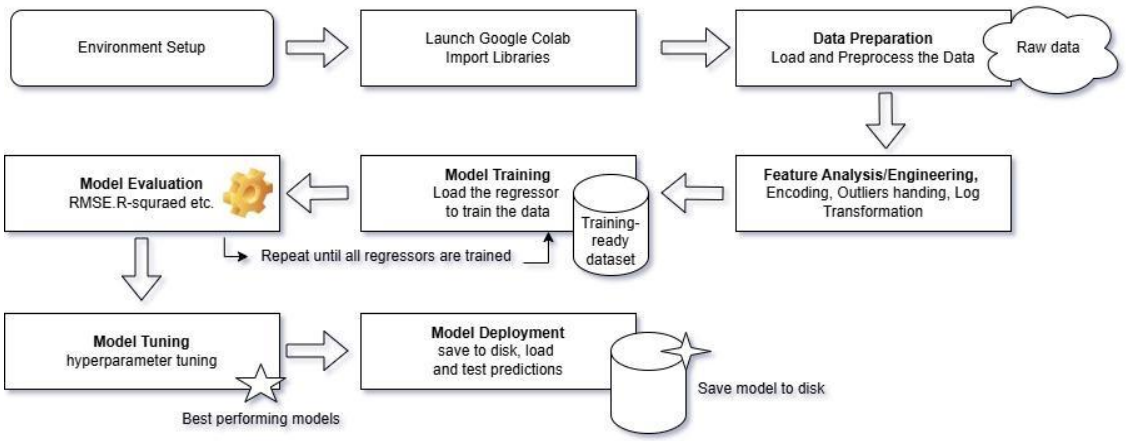

Fig. 3. A workflow diagram illustrating the core processes in the model development

https://doi.org/10.24191/mij.v6i2.9178

Table 2. Tools, platforms, and resources used in this study

Category	Details
Programming language	Python (using libraries such as Pandas, NumPy, Scikit-learn, TensorFlow, Keras)
Data Visualization	Matplotlib and Seaborn are used for exploratory data analysis (EDA), model evaluation, and presentation
Development Environment	Google Colab IDE (0.0.1a2) for implementation and experimentation; Visual Studio Code for local development
Cloud Resources	Google Colab IDE with CPU/GPU (T4) for training and deep learning
Operating System	Windows 11
Python Version	Python 3.12.11
Data Source	National Property Information Centre (NAPIC) - High-rise residential transactions (KL, 2021–2024)

4. RESULTS AND FINDINGS

This section presents the outcomes of the data analysis and model evaluation conducted to predict high-rise residential property prices in Kuala Lumpur using the CRISP-DM framework and a comprehensive dataset from the National Property Information Centre (NAPIC) spanning 2021–2024. Through exploratory data analysis (EDA), data preprocessing, feature engineering, and rigorous model comparison, this study evaluates the performance of multiple machine learning models—Multiple Linear Regression (MLR), Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machine (SVM), and Artificial Neural Network (ANN)—to identify the most accurate and interpretable approach. The findings highlight key predictors of property prices, model performance metrics, and practical implications for stakeholders in the real estate sector. The section also includes a real-world example of applying the model to a new condo and a summary of key results for quick reference. The section is organized into subsections covering EDA insights, data preprocessing and feature engineering, model performance evaluation, a real-world application example, model results summary, and model deployment results.

4.1 Exploratory Data Analysis (EDA)

EDA revealed critical insights into the dataset. The dataset, sourced from NAPIC (2021–2024), includes 12,735 high-rise residential transactions in Kuala Lumpur. Key features include land/parcel area (continuous), Mukim, Scheme Name/Area, Tenure, and Unit Level (categorical). Preprocessing involved: Univariate analysis, Fig. 4, showed skewed distributions for transaction prices and property areas, suggesting the need for transformations. Bivariate analysis indicated strong positive correlations between transaction price and features like Land/Parcel Area and Built-up Area. Categorical features such as Tenure and Mukim also showed distinct impacts on average transaction prices.

https://doi.org/10.24191/mij.v6i2.9178

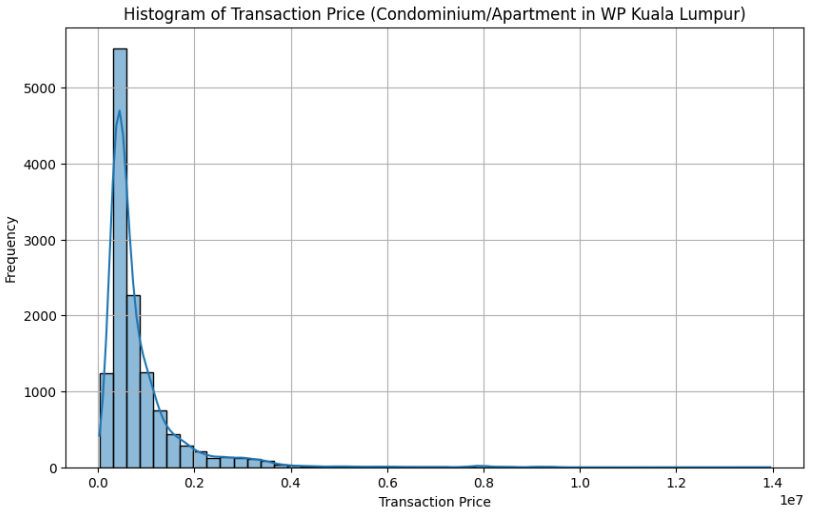

Fig. 4. Distribution of house transaction price

The line graph in Fig. 5 shows how the average and median transaction prices have changed over time monthly. It reveals the overall direction of the market (upward, downward, or stable). Market volatility (large month-to-month fluctuations) was observed from 2021-2022, potentially indicating market uncertainty or the influence of large, infrequent transactions during that period. The average is significantly higher than the Median, which suggests the presence of some very expensive properties skewing the average upwards. Overall, from 2022 to 2024, the first half, the market remains flat.

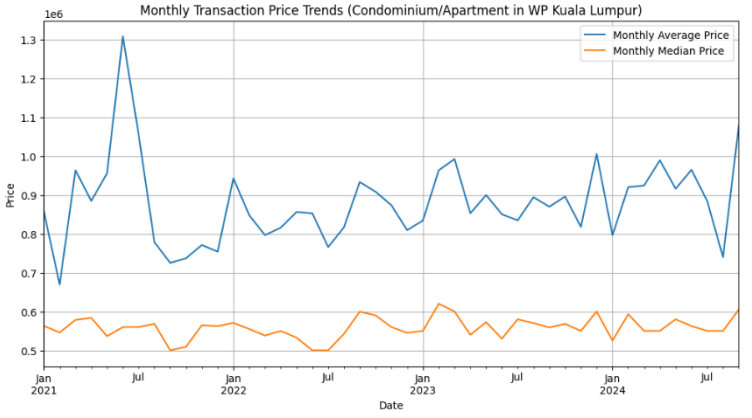

Fig. 5. Line graph showing monthly transaction price trends (average and median)

4.2 Data Preprocessing and Feature Engineering

Comprehensive data preprocessing and feature engineering were carried out to prepare the dataset for modelling. Initially, data cleaning involved the removal of duplicate entries and constant columns such as *Sector* and *District*, alongside other irrelevant features. Upon inspection, no missing values were detected, streamlining the preprocessing phase. Outliers were identified and removed using the Interquartile Range (IQR) method, which excluded 1,035 data points from *Transaction Price* (refer to Table 3) and 732 from *Land/Parcel Area* (see Table 4). Feature engineering efforts included applying a logarithmic transformation to reduce skewness in both *Transaction Price* (Fig. 6) and *Land/Parcel Area* (Fig. 7), and encoding categorical variables using label encoding techniques, detailed in Appendix A.1 to A.3.

https://doi.org/10.24191/mij.v6i2.9178

Table 3. Transaction price outliers' analysis

Description	Value
Lower Bound	--500,000.0
Upper Bound	1,900,000.0
Number of Outliers	1,035
Shape After Outlier Removal	(11,700, 7)

Table 4. Land/parcel area outliers' analysis

Description	Value
Lower Bound	180.5
Upper Bound	180.5
Number of Outliers	732
Shape After Outlier Removal	(10968, 7)

Log transformation was applied to skewed numerical features to approximate a normal distribution, improving model performance.

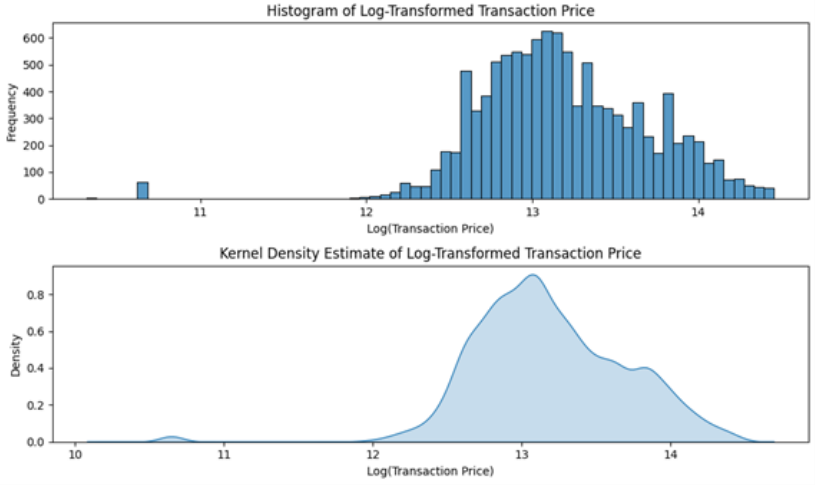

Fig. 6. Log Transformed transaction price distribution

https://doi.org/10.24191/mij.v6i2.9178

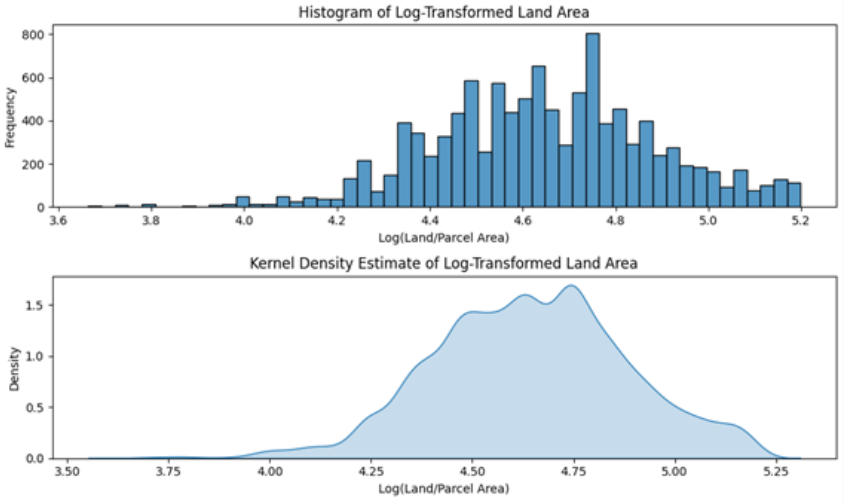

Fig. 7. Transformed land/parcel area distribution

Analysis of Feature importance (Fig. 8) highlighted Built-up Area, Land/Parcel Area, Tenure, and Mukim as the most significant predictors of house prices. Correlation analysis (Fig. 9) and feature importance rank (Fig. 10) (Table 5) identified Land/Parcel Area as the most influential predictor (r = 0.6446). Categorical features like Tenure and Mukim showed weaker correlations. The heatmap confirmed a strong positive correlation between Land/Parcel Area and Transaction Price (r = 0.6446).

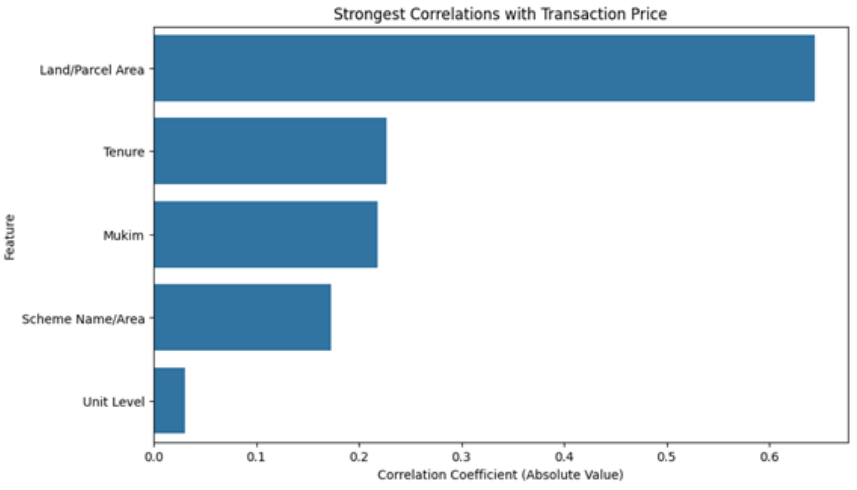

Fig. 8. Feature importance rank

https://doi.org/10.24191/mij.v6i2.9178

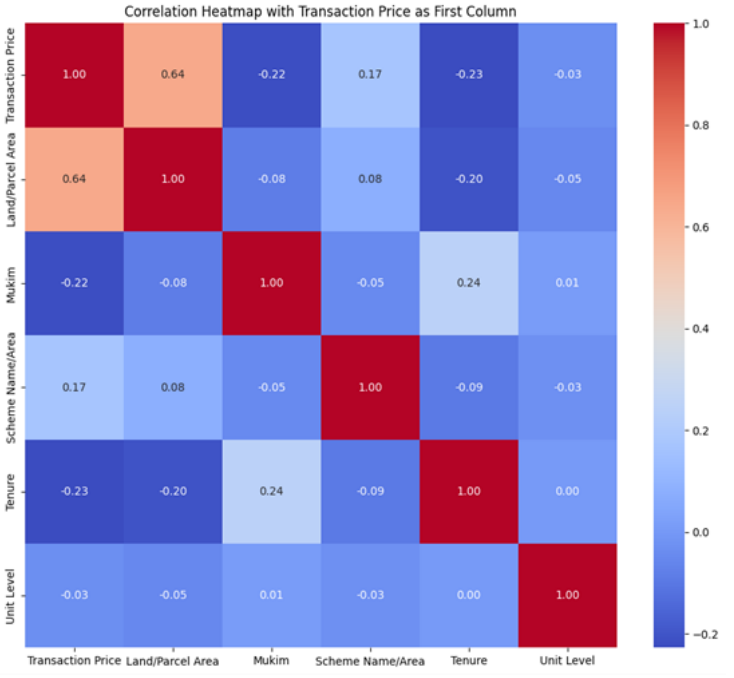

Fig. 9. Correlation heatmap

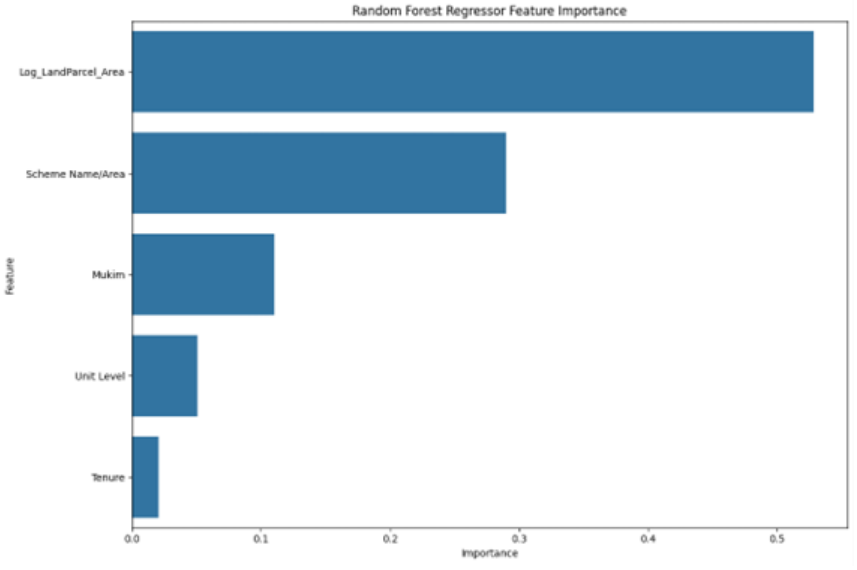

Fig. 10. Bar graph showing feature importance rank

Random Forest feature importance confirmed Log_LandParcel_Area as the dominant predictor (0.5281), followed by Scheme Name/Area (0.2901) Tenure had the lowest importance (0.0207) (Table 5).

https://doi.org/10.24191/mij.v6i2.9178

Table 5. Feature importance from Random Forest and Correlation Analysis.

Feature	Random Forest	Correlation Analysis
Log_LandParcel_Area	0.5281	0.6446
Scheme Name/Area	0.2901	0.1725
Mukim	0.1102	0.2179
Unit Level	0.0509	0.0308
Tenure	0.0207	0.2272

4.3 Model Performance

A comprehensive comparison of the developed models (MLR, DT, RF, GBM, SVM, ANN) demonstrated varying levels of predictive accuracy as shown in Table 6. Initial model runs indicated that ensemble methods generally outperformed single models. The Random Forest model's superior performance ($R^2 = 0.9025$) underscores its ability to capture non-linear relationships, outperforming linear models like MLR ($R^2 = 0.4423$). Land/parcel area's dominance aligns with prior studies (Abdullah & Mohd, 2022). The lower importance of categorical variables suggests that location and ownership type play secondary roles. Label encoding may have introduced ordinal biases, suggesting one-hot encoding for future work. Random Forest outperformed others, achieving the lowest MSE (0.0254), MAE (0.1033), RMSE (0.1592), and the highest R^2 (0.9025).

Table 6. Evaluation of metrics for baseline and advanced models

Model	MSE	R-squared	MAE	RMSE
Linear Regression	0.1451	0.4423	0.2821	0.3809
Decision Tree	0.0338	0.8702	0.1132	0.1838
Random Forest	0.0254	0.9025	0.1033	0.1592
GBM	0.0595	0.7712	0.1756	0.2440
SVM	0.0602	0.7686	0.1643	0.2453
ANN	0.1103	0.5759	0.2389	0.3321

K-fold cross-validation and hyperparameter tuning enhanced robustness. The feature importance analysis from the optimal Random Forest model confirmed the critical influence of Built-up Area, Land/Parcel Area, and Tenure on predicted prices, providing valuable insights into the key drivers of the Kuala Lumpur high-rise residential market. Following K-fold cross-validation and hyperparameter tuning, improvements in model stability and generalization were observed. The Random Forest model consistently achieved the lowest error rates (MAE, MSE, RMSE) and the highest R^2 score, indicating its superior predictive power and ability to explain the variance in house prices as shown in Table 7. Gradient Boosting Machine also performed commendably, while Artificial Neural Networks showed promise but required more extensive tuning for optimal performance. After tuning, Random Forest improved slightly (Table 8) with no significant overfitting (Table 9).

Table 7. Evaluation metrics for baseline and advanced models after applying K-Fold validation

Model	MSE	R-squared	MAE	RMSE
Linear Regression	0.146695	0.440019	0.281606	0.383008
Decision Tree	0.038210	0.85428	0.114247	0.195474
Random Forest	**0.027413**	**0.895305**	**0.103500**	**0.165569**
GBM	0.061774	0.763999	0.177108	0.248544
SVM	0.067220	0.743143	0.165594	0.259174
ANN	0.130978	0.500152	0.259620	0.361729

https://doi.org/10.24191/mij.v6i2.9178

Table 8. Random Forest performance before and after hyperparameter tuning.

Evaluation Metric	Before Tuning	After Tuning	Difference
Mean Squared Error	0.0274	0.0257	-0.0017
R-squared	0.8953	0.9013	+0.0060

Table 9. Random Forest performance on training and testing data

Evaluation Metric	Training Value	Testing Value
Mean Squared Error	0.0103	0.0257
R-squared	0.9607	0.9013

4.4 Model Deployment

The tuned Random Forest model predicted a property price of RM 472,222 for a leasehold unit and RM 601,686 for a freehold unit, highlighting a RM 129,464 premium for freehold properties (Table 10).

Table 10. Sample input variables and predicted prices

Variable Type	Variable Name	Encoded/Transformed Value	Original Value
Categorical	Mukim	2	Mukim Batu
Categorical	Scheme Name/Area	69	CASA IDAMAN
Categorical	Tenure	1	Leasehold
Categorical	Unit Level	55	7
Continuous	Log_LandParcel_Area	4.70953	111 square meters

Predicted Log Price: 13.0652, Price: RM 472,222 (Leasehold)

Predicted Log Price: 13.3075, Price: RM 601,686 (Freehold)

4.5 Real-World Application Example

To demonstrate the practical utility of the Random Forest model, we applied it to predict the price of a hypothetical new condominium unit in Kuala Lumpur's Mukim Setapak region. The condo has a land/parcel area of 120 square meters, is in a new development (encoded as Scheme Name/Area 419), has a freehold tenure (encoded as 0), and is on the 22nd floor (encoded as Unit Level 17). Using the tuned Random Forest model, the predicted price is RM 705,386. This example illustrates how the model can assist stakeholders, such as developers pricing new units or homebuyers evaluating investment options, by providing accurate and data-driven price estimates tailored to specific property characteristics.

4.6 Model Results Summary

For quick reference, the key results of the model evaluation are summarized in Table 11 below, highlighting the superior performance of the Random Forest model and the most influential predictors.

Table 11. Summary of Key Model Performance and Feature Importance

Category	Detail
Best Model	Random Forest (R^2 = 0.9025, MSE = 0.0254, MAE = 0.1033, RMSE = 0.1592)
Other Models	Decision Tree (R^2 = 0.8702), GBM (R^2 = 0.7712), SVM (R^2 = 0.7686), ANN (R^2 = 0.5759), MLR (R^2 = 0.4423)
Key Predictors	Log_LandParcel_Area (RF Importance: 0.5281, Correlation: 0.6446), Scheme Name/Area (0.2901), Tenure (0.0207)
Practical Insight	Property size and the characteristics of the project or scheme are the main determinants of property pricing. Freehold properties command a RM 129,464 premium over leasehold (e.g., RM 601,686 vs. RM 472,222 for a 111 m^2 unit)

https://doi.org/10.24191/mij.v6i2.9178

5. CONCLUSION AND RECOMMENDATIONS

This study successfully developed and evaluated a data-driven machine learning framework for predicting high-rise residential property prices in Kuala Lumpur. By leveraging real-world NAPIC data and applying the CRISP-DM methodology, the research demonstrated the superior predictive capabilities of advanced machine learning models over traditional approaches. The Random Forest algorithm was identified as the optimal model, offering high accuracy and interpretability. Random Forest offers a robust approach to predicting high-rise residential property prices in Kuala Lumpur. Land/parcel area was the primary driver, with Random Forest achieving high accuracy (R^2 = 0.9025). The Random Forest model's superior performance (R^2 = 0.9025) underscores its ability to capture non-linear relationships, outperforming linear models like MLR (R^2 = 0.4423). Land/parcel area's dominance aligns with prior studies (Abdullah & Mohd, 2022). The lower importance of categorical variables suggests that location and ownership type play secondary roles. Label encoding may have introduced ordinal biases, suggesting one-hot encoding for future work. K-fold cross-validation and hyperparameter tuning enhanced robustness.

The study provides valuable insights into the significant features influencing house prices, such as built-up area, land area, and tenure, which are crucial for informed decision-making by various stakeholders in the real estate sector. This research contributes to the existing body of knowledge by providing a systematic comparison of multiple machine learning models on a comprehensive Malaysian housing dataset, addressing the existing research gap in applying advanced ML techniques like ANNs in this context. The identified optimal model and key influencing factors offer practical guidance for investors, homebuyers, and urban planners. These findings support stakeholders in making data-driven decisions. However, the current model has limitations, including the lack of external validation on datasets from other regions, which may limit generalizability, and the use of label encoding for categorical variables, which may introduce ordinal biases. These limitations highlight the need for further refinement to enhance model robustness and applicability.

Future research could explore the integration of geospatial data (e.g., proximity to amenities, public transport) to enhance prediction accuracy. Investigating advanced deep learning architectures, such as Recurrent Neural Networks (RNNs) for time-series forecasting, or experimenting with transfer learning from other property markets, could further improve model performance. To address the scalability limitation noted in the Literature Review, future research should propose testing the model's generalizability on high-rise residential property data from other Malaysian cities, such as Penang and Johor Bahru, or international markets to evaluate its predictive performance across diverse real estate contexts. Additionally, developing a user-friendly web application for real-time predictions based on the deployed model would increase its practical utility.

6. ACKNOWLEDGEMENTS/FUNDING

The author expresses sincere appreciation to their main thesis supervisor, Dr. Doris Wong Hooi Ten, and co-supervisor Assoc. Prof. Ts. Dr. Maslin Masrom, for their invaluable encouragement, guidance, and unwavering support throughout this research. The author also acknowledges Universiti Teknologi Malaysia (UTM) for providing funding and resources, as well as the UTM librarians for their assistance. Special thanks are extended to fellow postgraduate students, colleagues, and family for their continuous support and understanding.

7. CONFLICT OF INTEREST STATEMENT

The authors agree that this research was conducted in the absence of any self-benefits, commercial or financial conflicts and declare the absence of conflicting interests with the funders.

https://doi.org/10.24191/mij.v6i2.9178

8. AUTHORS' CONTRIBUTIONS

Lim Eng Lian conceptualized and designed the study, was responsible for data acquisition and curation, performed the data analysis and model development, conducted the investigation, and prepared the original draft of the manuscript. The author has read and agreed to the published version of the manuscript.

REFERENCES

Bau, Y., & Hisham, S. M. S. B. (2022). A case study using machine learning techniques for prediction of house prices in WP, Malaysia. In *Proceedings of the International Conference on Computer, Information Technology and Intelligent Computing (CITIC 2022)* (pp. 79–91). https://doi.org/10.2991/978-94-6463-094-7_7

Chan, F., Schulz, R., & Zhang, Z. (2023). An application of machine learning in real estate economics: What extra benefits could machine learning techniques provide? In J. Vaze, C. Chilcott, L. Hutley, & S. M. Cuddy (Eds.), *Proceedings of the 25th International Congress on Modelling and Simulation, MODSIM 2023* (pp. 123–129). Modelling and Simulation Society of Australia and New Zealand Inc. https://doi.org/10.36334/modsim.2023.chan39

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chun, H. J., Lee, U. H., & Lee, B. G. (2025). Predicting housing price in Seoul using explainable AI (XAI) and machine learning. *KSII Transactions on Internet and Information Systems, 19*(4), 1077–1096. https://doi.org/10.3837/tiis.2025.04.002

Gulati, V., & Raheja, N. (2021). Efficiency enhancement of machine learning approaches through the impact of preprocessing techniques. In *Proceedings of the 6th International Conference on Signal Processing, Computing and Control (ISPCC)* (pp. 1–6). IEEE. https://doi.org/10.1109/ISPCC53510.2021.9609474

Ho, W. K., Tang, B., & Wong, S. W. (2020). Predicting property prices with machine learning algorithms. *Journal of Property Research, 38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Ja'afar, N. S., Mohamad, J., & Ismail, S. (2021). Machine learning for property price prediction and price valuation: A systematic literature review. *Planning Malaysia, 19*. https://doi.org/10.21837/pm.v19i17.1018

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. A. (2021). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering, 33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

McCluskey, W., Zulkarnain Daud, D., & Kamarudin, N. (2014). Boosted regression trees. *Journal of Financial Management of Property and Construction, 19*(2), 152–167. https://doi.org/10.1108/JFMPC-06-2013-0022

Rahmat, F., Zulkafli, Z., Ishak, A. J., Rahman, R. Z. A., De Stercke, S., Buytaert, W., Tahir, W., Rahman, J. A., Ibrahim, S., & Ismail, M. (2023). Supervised feature selection using principal component analysis. *Knowledge and Information Systems, 66*(3), 1955–1995. https://doi.org/10.1007/s10115-023-01993-5

Rattanaprichavej, N., & Teeramungcalanon, M. (2020). An investment decision: Expected and earned yields for passive income real estate investors. *Cogent Business & Management, 7*(1).

https://doi.org/10.24191/mij.v6i2.9178

https://doi.org/10.1080/23311975.2020.1786331

Ravikumar, M., & Nagashree, J. (2023). Enhancing predictive modeling in Kuala Lumpur real estate: A comprehensive data preprocessing and feature engineering approach. *International Journal of All Research Education and Scientific Methods, 12*(04), 833–838. https://doi.org/10.56025/ijaresm.2023.120124833

Sa'at, N. F., Maimun, N. H. A., & Idris, N. H. (2021). Enhancing the accuracy of Malaysian house price forecasting: A comparative analysis on the forecasting performance between the hedonic price model and artificial neural network model. *Planning Malaysia, 19*(17). https://doi.org/10.21837/pm.v19i17.1003

StarProperty. (2024). Malaysia's property market: A transformative outlook for 2025. https://www.starproperty.my/news/malaysia-s-property-market-a-transformative-outlook-for-2025/130769

Tahir, T., & Fatima, G. (2024). Exploring data-driven real estate price prediction in a developing country: The case of Pakistan. In *2024 International Conference on Frontiers of Information Technology (FIT)* (pp. 1–5). IEEE. https://doi.org/10.1109/fit63703.2024.10838414

Ved, A. R., & Gupta, A. (2024). Extensive data preprocessing and exploratory data analysis for house price prediction using machine learning and deep learning with an introduction to AutoML technologies. In *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC 2024)* (pp. 711–721). https://doi.org/10.1109/AIC61668.2024.10730866

Wang, X., Gao, S., Zhou, S., Guo, Y., Duan, Y., & Wu, D. (2021). Prediction of house price index based on bagging integrated WOA–SVR model. *Mathematical Problems in Engineering, 2021*, 1–15. https://doi.org/10.1155/2021/3744320

https://doi.org/10.24191/mij.v6i2.9178

APPENDIX

A. Label Encoding Tables

A.1. Mukim - encoded with 8 categories representing Kuala Lumpur's administrative regions

Original Value	Encoded Value
Kuala Lumpur Town Centre	0
Mukim Ampang	1
Mukim Batu	2
Mukim Cheras	3
Mukim Kuala Lumpur	4
Mukim Petaling	5
Mukim Setapak	6
Mukim Ulu Kelang	7

A.2. Tenure

Original Value	Encoded Value
Freehold	0
Leasehold	1

A.3. Unit Level (only a subset of the full encoding scheme is shown here for brevity)

Original Value	Encoded Value
03A	0
1	1
10	2
…	…
…	…
G	59
LG	60
MZ	61
UG	62

https://doi.org/10.24191/mij.v6i2.9178