UNIVERSITI MALAYA

# WQD 7006

# Machine Learning For Data Science

# Case Study

# Property Price Prediction using Machine Learning Algorithms

**Lecturer: Dr. Muhammad Shahreeza Safiruz Bin Kassim**

**Group Members:**

| Name | ID |
|---|---|
| Nur Hidayah binti Ahmad Shafii | 22120931 |
| Choon Yue Hua | 17152027 |
| Then Dao Qing | 23057608 |
| Low Meng Fei | 23063305 |
| Choo Wan Qi | 23069166 |

## Introduction

In Malaysia, the demand for residential properties continues to grow steadily. This growth is driven by urbanization, promotional efforts by developers, job opportunities and well-developed infrastructure. This surge in demand highlights the critical need for accurate price prediction models. Rising housing costs have made homeownership increasingly challenging. Therefore, it is important to estimate the possible value of a property to evaluate market value and identify if the property is over-pricing or under-pricing. Understanding the rental price dynamics not only helps manage market value but also directly impacts the financial decisions of buyers and sellers. With the advancement of machine learning algorithms, the ability to predict property prices has been significantly improved. By analyzing large datasets, these algorithms can identify trends, forecast future prices, and improve decision-making processes in the real estate market.

## Research Gaps

The research gap identified in the studies highlight several areas where further investigation is required to improve the application of machine learning in real estate valuation. In the study on housing price prediction in Petaling Jaya, it was mentioned that the application of machine learning algorithms in the Malaysian real estate market remains limited. In particular, there is a lack of research focused on localized housing markets in urban areas like Petaling Jaya. This indicates a need for more targeted research in this context. The study’s findings are based on a specific dataset limits their generalizability to other markets or scenarios. Additionally, the study emphasizes that irrelevant features in datasets can reduce the accuracy of predictive models. To address these gaps, further investigation is needed to evaluate the impact of dataset characteristics, parameter tuning, and feature selection on the accuracy of machine learning models for housing price prediction. Moreover, researchers should explore different types of housing price prediction problems in Malaysia to provide broader insights and improve the applicability of machine learning in this field.

In the research on heritage property valuation using machine learning, we found quite similar gaps whereby there is lack of application of machine learning techniques in Malaysia's real estate market, particularly in terms of heritage properties valuation. Furthermore, there is no established standard to identify the historical characteristics that are important for heritage property valuation. The dataset used in the study was relatively small and lacked sufficient location-specific and historical details. While traditional methods like regression models have been explored, there is limited comparison with machine learning techniques like Neural Networks or Random Forests in heritage property contexts. Also, the non-linear market characteristics were not fully addressed in the study. Hence, future research should incorporate detailed historical and location-specific attributes relevant to heritage properties. Comparative studies between traditional methods and advanced machine learning approaches should be done to determine the strengths and limitations. Lastly, future investigation on the impact of non-linear market dynamics using advanced algorithms could improve model generalizability.

## Methodology and Results

Both journals utilised 5 models for data modelling, Ja’afar et al. used Neural Network, Random Forest, Support Vector Machine, K-Nearest Neighbors and Linear Regression. On the other hand, Masrom et al. used Linear Regression, Random Forest, Decision Tree, Lasso and Ridge. The Neural Network was implemented with weight decay to avoid overfitting problems and Random Forest was selected because the dependent variable in this study is continuous value which applies regression tree. Support Vector Machine modified the cost function for distance measuring and KNN standardizes the variable to eliminate the difference in scale to cater continuous variables. Linear Regression utilised linear combinations of independent variables to estimate a continuous dependent variable. Additionally, Lasso was employed to remove irrelevant features by shrinking the coefficient to zero, while Ridge was used to resolve inaccurate estimation of prediction due to multi-co-linearity in the dataset.

Datasets used on journal from Ja’afar et al. is originated from NAPIC which has the data of prewar shophouse in Penang Island, Malaysia from 2004 to 2018. This dataset has 3121 number of records reduced to 248 which excludes the incompleteness data and data that is not in the designated area. While the journal from Masrom et al did not mention any info about the dataset sources.

Moving on to the features of datasets, there are 6 features in the datasets of Ja’afar et al. journal which are Price, Road, Zone, Storey, Year and Lot Size. While, Masrom et al. journal datasets have 15 features which are Buying Price, Floor, Green certificate, Main floor area, Number of bedrooms, Distance of CBD, Building category, Ownership, Category area, Area classification, Floor, Build classification, Age of building, buyers and seller. Both journals show weak or very weak correlations among the features except for the buying price and selling price features in Masrom et al. Journal shows strong correlation.

Then, on the evaluation part, Ja’afar et al. journal uses 3 evaluation metrics which are Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R squared ($R^2$) while Masrom et al. journal only uses RMSE as the evaluation metric. In summary both journals show that Random Forest is the best algorithm to use in predicting property price with high $R^2$ and low error in prediction.

### Strength and Weakness

| | Masrom et al | Ja’afar et al |
|---|---|---|
| Strength | 1. **Focused Feature Selection**: By grouping the data according to correlation strength (strong, moderate, weak, very weak) helps the analysis target the impact of features that are directly related to selling price.<br>2. **Provide results by testing all models across all groups (All, Strong, Moderate, Weak, and Very Weak)**: This helps reveal how predictive power changes when weaker correlations are included or excluded, which can help in optimizing feature engineering. | 1. **More Evaluation Metrics- $R^2$, RMSE, and MAE:** Offers a more comprehensive understanding of the model's performance across several dimensions. |
| Weakness | 1. **Potential Bias in Grouping:** Uneven ranges for "strong," "moderate," and "weak" correlations might affect the generalizability of results.<br>**Strong**: 0.51 to 1 (a large range)<br>**Moderate**: 0.3 to 0.5 (a smaller range)<br>**Weak**: 0.2 to 0.29 (very tight)<br>**Very Weak**: 0.1 to 0.19 (very tight)<br>2. **Correlation does not capture non-linear relationships:** Potentially useful features may be overlooked or misclassified as weak.<br>3. **Limited coverage:** Focus restricted to Petaling Jaya, Selangor | 1. **Does not address the non-linearity:** Could potentially improve model accuracy when non-linear models are used.<br>2. **Limited coverage:** Focusing only on the core zone and buffer zone for UNESCO World Heritage Sites at North Eastern tip of Penang<br>3. **Lesser features used:** Only includes 6 features in total - price, road, zone, storey, year and lot size. |

## Recommendations

| Improvements | Reasons |
|---|---|
| Data Collection and Preprocessing<br>● Expand geographical coverage by including more areas and property types<br>● Include property condition assessments<br>● Collect temporal data spanning longer periods (5-10 years) | ● Improve model generalizability<br>● Better reflect real estate market dynamics<br>● Account for location value drivers |
| Feature Engineering and Selection<br>● Implement advanced feature selection methods such as PCA and RFE<br>● Create interaction terms between key features<br>● Develop composite features such as price per square foot | ● Able to capture complex relationships between features<br>● Improve model accuracy and interpretability<br>● Identify truly significant predictors |
| Algorithm Enhancement<br>● Use advanced algorithms like XGBoost and LightGBM<br>● Perform extensive hyperparameter tuning using grid search with cross-validation<br>● Implement AutoML pipelines | ● Improve prediction accuracy<br>● Reduce overfitting<br>● Automate model selection and tuning |
| Evaluation Framework<br>● Implement k-fold cross-validation<br>● Conduct feature importance analysis<br>● Perform model robustness tests<br>● Implement confidence intervals for predictions | ● Provide more reliable performance assessment<br>● Better understand model limitations<br>● Identify areas for improvement |
| Practical Deployment<br>● Develop API endpoints for model deployment<br>● Implement monitoring systems for model performance<br>● Develop user-friendly interfaces<br>● Include model explanation tools such as SHAP values | ● Make the model practically useful<br>● Ensure sustainable model performance<br>● Improve user adoption<br>● Provide transparency in predictions |

## Reflection

In this project, we explore the use of machine learning models in property valuation. Masrom et al. (2019) investigated the use of machine learning models in valuation of residential housing prices in Petaling Jaya, Selangor, which is a general investigation topic. This topic is practical and its output could provide insights to the public as well as authorities. Meanwhile,  Ja'afar et al. (2020) focuses on the valuation of heritage buildings in Penang, specifically the pre-war shophouse prices in Malaysia. This study focuses on the underexplored area, since no one has ever done this study before and this could provide a fresh perspective on the potential of  machine learning in predicting heritage building valuation.

In terms of the methodology, Masrom et al. (2019) emphasized feature selection impact with studies conducted using different combinations of features. The authors found that using the strong correlation features alone could achieve similar performance to using all features, and hence, the result suggests the use of strong and moderately correlated variables to generate the most accurate and reliable prediction. The result also showed that buying price has a high correlation with selling price compared to other features such as floor area.  Ja'afar et al. (2020) focuses on algorithm comparison. In terms of the datasets, Masrom et al. (2019) included more features (16), while Ja'afar et al. (2020) only focused on 6 features. This is due to the unique characteristics of heritage building, the data is limited and more challenging for data collection. Besides, heritage buildings' values lie in its historical and cultural attributes, which is especially hard to capture in terms of variables.

In terms of the result, both papers compare 5 machine learning algorithms. Masrom et al. (2019) compared Linear Regression, Random Forest, Decision Tree, Lasso and Ridge, meanwhile Ja'afar et al. (2020) compared Neural Networks, Random Forest,  Support Vector Machine, k-Nearest Neighbors and Linear Regression. Both found Random Forest to be the best performing algorithm, capable of handling different types of data. Similar validation metrics are used such as $R^2$ and RMSE. The $R^2$ value of both Random Forest models are close to 1, indicating that the result produced is highly following the actual trend.

In conclusion, both studies contribute to the property price prediction in Malaysia in different locations and property types. The use of machine learning models could provide reliable prediction in the real estate market. This prediction is useful and could be used for implementations such as providing data for the authority to do urban development decisions or for banking institutions to assess property values for mortgage lending.

**References (papers used)**

1) Mohamad, J., Ja'afar, S., & Ismail, S. (2020). Heritage Property Valuation using Machine Learning Algorithms. Annual Pacific Rim Real Estate Society, 1–12.
2) Mohd, T., Masrom, S., & Johari, N. (2019). Machine Learning Housing Price Prediction in Selangor, Malaysia. Recent Technology and Engineering (RTE), 8(2), 542–546.