LOW JUN LIANG

B.Sc. (Hons) Statistical Computing and Operations Research 2025

ENHANCING HOUSE PRICE PREDICTION USING HYBRID FEATURE SELECTION: A COMBINATION OF INFORMATION GAIN AND SVM-RFE

LOW JUN LIANG

BACHELOR OF SCIENCE (HONOURS) STATISTICAL COMPUTING AND OPERATIONS RESEARCH

FACULTY OF SCIENCE

UNIVERSITY TUNKU ABDUL RAHMAN

MAY 2025

ENHANCING HOUSE PRICE PREDICTION USING HYBRID FEATURE SELECTION: A COMBINATION OF INFORMATION GAIN AND SVM-RFE

By

LOW JUN LIANG

A project report submitted to the

Department of Physical and Mathematical Science

Faculty of Science

Universiti Tunku Abdul Rahman

in partial fulfilment of the requirements for the degree of

Bachelor of Science (Honours)

Statistical Computing and Operations Research

May 2025

ABSTRACT

ENHANCING HOUSE PRICE PREDICTION USING HYBRID FEATURE SELECTION: A COMBINATION OF INFORMATION GAIN AND SVM-RFE

LOW JUN LIANG

Accurate house price prediction is crucial for buyers, investors, and policymakers to make informed decisions. However, real estate datasets often contain high-dimensional features, including redundant and irrelevant attributes, which can negatively impact model performance. This study proposes a hybrid feature selection approach that combines Information Gain (IG) and Support Vector Machine Recursive Feature Elimination to enhance predictive accuracy. The proposed hybrid method significantly improves model performance, achieving a 22.2% reduction in Root Mean Squared Error (RMSE) (from 185,518.52 to 154,403.70) and a 22.7% increase in R-squared (from 0.6522 to 0.8008) compared to using IG alone. While IG is effective in ranking features based on their relevance to the target variable, it does not account for feature interactions and redundancy, which can lead to suboptimal feature selection. The addition of SVM-RFE addresses this limitation by iteratively refining the feature set, ensuring only the most informative attributes are retained. Furthermore, the hybrid approach demonstrated robustness even in the presence of artificially introduced noise. Hyperparameter tuning further optimized the best-performing model, yielding

marginal improvements in accuracy. These findings highlight the effectiveness of combining filter and wrapper methods for real estate price prediction, demonstrating that hybrid feature selection leads to more reliable and interpretable models.

ACKNOWLEDGEMENTS

I am very happy to undertake the final year project entitled “Enhancing House Price Prediction Using Hybrid Feature Selection: A Combination of Information Gain and SVM-RFE”. I would like to take this opportunity to thank every individual involved in this project. High appreciation is given to my supervisor, Dr. Chin Fung Yuen, a lecturer at the Faculty of Science at Universiti Tunku Abdul Rahman, for her strong and timely support in developing my project and writing this report. Most importantly, I am grateful to Universiti Tunku Abdul Rahman, for giving me a chance to study for the program Bachelor of Science (Hons) in Statistical Computing and Operations Research. Lastly, special thanks to my family members and friends who encourage me to complete this project.

`

DECLARATION

I hereby declare that the project report is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTAR or other institutions.

LOW JUN LIANG

APPROVAL SHEET

This project report entitled "**ENHANCING HOUSE PRICE PREDICTION USING HYBRID FEATURE SELECTION: A COMBINATION OF INFORMATION GAIN AND SVM-RFE**" was prepared by LOW JUN LIANG (ID No: **22ADB02353**) and submitted as partial fulfilment of the requirements for the degree of Bachelor of Science (Hons) Statistical Computing and Operations Research at Universiti Tunku Abdul Rahman.

Approved by:

(Dr. Chin Fung Yuen) Date: **13 April 2025**

Supervisor

Department of Physical and Mathematical Science

Faculty of Science

Universiti Tunku Abdul Rahman

FACULTY OF SCIENCE

UNIVERSITI TUNKU ABDUL RAHMAN

Date: 9 April 2025

PERMISSION SHEET

It is hereby certified that **LOW JUN LIANG** (ID No: **22ADB02353**) has completed this final year project entitled "**ENHANCING HOUSE PRICE PREDICTION USING HYBRID FEATURE SELECTION: A COMBINATION OF INFORMATION GAIN AND SVM-RFE**" under the supervision of Ms. Chin Fung Yuen (Supervisor) from the Department of Physical and Mathematical Science, Faculty of Science.

I hereby give permission to the University to upload the softcopy of my final year project in pdf format into the UTAR Institutional Repository, which may be made accessible to the UTAR community and public.

Yours truly,

(LOW JUN LIANG)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

IG	Information Gain
JPPH	Valuation and Property Services Department Malaysia
LightGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic explanations
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MdaPE	Median Absolute Percentage Error
MI	Mutual Information
MRT	Mass Rapid Transit
MSE	Mean Square Error
OLS	Ordinary Least Squares
RMSE	Root Mean Squared Error
RMSLE	Root Mean Squared Logarithmic Error
SHAP	Shapley Additive Explanations
SVR	Support Vector Regression

SVM	Support Vector Machine
RFE	Recursive Feature Elimination
XGBoost	eXtreme Gradient Boost

CHAPTER 1

INTRODUCTION

1.1 Background of Study

An accurate prediction of housing prices is a cornerstone for various stakeholders, such as investors, policymakers, and homeowners, who rely on correct pricing to make sound judgments in the housing market. Accurate price predictions reduce financial risks and prevent overpricing or underpricing, which is a gain for buyers and investors (Wang et al., 2019).

Past methods, like linear regression and hedonic price models, have also commonly been used to predict housing prices. These, however, prove to be inappropriate because they are not able to account for the sophisticated way that different housing features impact market price. They frequently rely on data distribution assumptions that may not reflect real-world behavior (Liang and Yuan, 2021). This limits them in scenarios where one works with big and complex data sets consisting of a high number of interconnected and noisy variables (Wei et al., 2022).

With the advent of the internet, a vast amount of information is available about properties. Internet portals such as Mudah in Malaysia provide abundant amounts of information on property listings in terms of location, price, area, and amenities.

These kinds of datasets are goldmines for researchers working on trends and patterns in housing markets (Sharma et al., 2024).

In recent years, advanced machine learning techniques have gained more popularity as property price forecasting tools (Mathotaarachchi et al., 2024; Ho et al., 2020). The techniques tackle issues better than the traditional method by finding complex patterns in large data sets, provided that they are provided with quality data. One of the most important steps in that direction is feature selection which is selecting the most relevant factors and eliminating unnecessary ones (Wang & Xu, 2019).

Feature selection can be grouped under three general classes which are filter methods, wrapper methods, and embedded methods. These methods address datasets in different ways towards modeling (Kumar & Minz, 2020). Models become interpretable with less overfitting by using suitable feature selection methods, thus achieving better predictive outcomes. Hybrid approaches, which bring together various feature selection methods, also provide encouraging outcomes. One of the top-ranked filter algorithms is Information Gain (IG), which orders features based on similarity to the target variable. IG does not take into consideration inter-feature interdependence and hence may result in duplicated selections. Alternatively, Support Vector Machine Recursive Feature Elimination (SVM-RFE) is a wrapper that recursively eliminates irrelevant features but is computationally costly, especially for big data sets. This study will try to highlight

a hybrid strategy of combining SVM-RFE and IG to determine the most relevant features and improve house price predictions.

1.2 Problem Statement

Housing price prediction is a difficult endeavor because real estate data is very complex. Many factors, such as local data, economic factors, and some individualistic properties of buildings, can mask useful tendencies. The location or size of a property is very important in determining the price, while other features like pools or gardens may be less important or may vary in importance depending on the given situation (Chen et al., 2020). Apart from this, interactions among these variables may be non-linear and complicated, which makes it difficult for traditional techniques to handle adequately.

Apart from that, traditional methods are tested in handling the non-linear and complex interdependence among such variables. Underfitting is caused by poor feature selection that hides useful patterns in redundant data, alongside overfitting potential, whereby models generalize well to training data but not when applied (Guo, 2023). Incorrect predictions reduce trustworthiness, particularly in critical applications where trustworthy models are critical, such as banking and government planning.

Traditional feature selection methods struggle to optimize accuracy and efficiency when applied to such complex, high-dimensional data (Yazdani, 2021). Simple filtering methods like Information Gain (IG) are susceptible to quickly score individual features in terms of importance but ignore potential interactions and

redundancy among features. Conversely, computationally exhaustive wrapper methods such as Support Vector Machine Recursive Feature Elimination perform well at picking the best feature subsets but are slow when handling big data. These inadequacies underscore the need for advanced feature selection techniques that integrate the strengths of both methodologies and therefore enhance model performance and offer precise house price approximations (Kostic & Jevremovic, 2021).

1.3 Objectives of Research

The research objectives for this project are:

1. To eliminate redundant or irrelevant features and determine the most significant features affecting house price using the hybrid of IG and SVM-RFE, thereby minimizing noise and improving the effectiveness and accuracy of predictive models.

2. To evaluate the hybrid method's performance against traditional feature selection techniques, such as filter by focusing on prediction accuracy and model robustness.

3. To examine whether and how hyperparameter optimization of the prediction model, performed after using the hybrid feature selection technique, improves house price prediction accuracy.

1.4 Significance of Study

This study is crucial for both learning and practical uses. It deals with a major issue in large real estate datasets, which is choosing the right features to accurately predict house prices. The study proposes a method that blends Support Vector Machine Recursive Feature Elimination and Information Gain (IG). The method reinforces predictability by removing redundant information and allowing one to better understand related feature dynamics than with the conventional methodologies. The study can potentially become a benchmark in its use of machine learning in property studies.

Third, the findings of the study can impact many stakeholders. Investors may employ sophisticated price-predicting models to spot undervalued properties, obtain portfolio diversification, and reduce financial risks (Shi et al., 2021). Policy makers and government regulators may employ the information to formulate equitable housing policy, construct suitable taxation frameworks, and predict market patterns, thus creating a more stable housing market (Vargas-Calderón & Camargo, 2022). The homebuyers can be assisted by the provision of more accurate and transparent valuations of property, which may assist them in making highly educated home purchasing decisions and avoid overpayment or underpayment for houses (Wang et al., 2019). This erases the potential risks of overcharging or undercharging, indicative of the usefulness of the study in economic decision-making.

Lastly, this research contributes to the field of machine learning by demonstrating the efficiency of hybrid feature selection in dealing with complex data. Its findings can be extended to other areas like marketing, finance, or medicine, where appropriate features need to be chosen and dealt with high-dimensional data are also equally significant. This study shows the way in which predictions can be enhanced using hybrid methods, presenting new understanding and application to real issues, and opening the door for further research in data-driven decision-making.

CHAPTER 2

LITERATURE REVIEW

The application of machine learning in predicting residential property values is gaining more traction, as it makes accurate forecasts of home prices in the future that are helpful to a diverse group of stakeholders. Various machine learning methods, including linear regression, decision trees, and neural networks, have been employed in scholarly research in housing price prediction.

2.1 Evaluation of Machine Learning Models

Abdul-Rahman et al. (2021) examine four single machine learning models employed in predicting house property prices, which include Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Ridge Regression, and Multiple Regression Analysis (MRA). The researchers employed Kuala Lumpur data and performed a comparison of model performance based on mean absolute error (MAE) and root mean square error (RMSE). Among the models examined, XGBoost demonstrated greater accuracy. This means that machine learning can indeed improve price forecasting in Malaysia to the advantage of policymakers, investors, and homeowners.

Yee et al. (2021) employ a dataset from the Malaysian Valuation and Property Services Department to forecast residential property prices using three machine learning models, which are Random Forest, Decision Tree, and Linear Regression. The models were validated with performance measures such as accuracy, R-squared (R^2), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Random Forest worked best with good accuracy. Decision Tree was very accurate but performed badly with middle-range prices, while Linear Regression was worse with complex data patterns. This indicates that Random Forest can give great insight when used to predict the prices of properties.

Furia and Khandare (2022) discuss the application of machine learning methods for property price prediction. The paper discusses the usage of deep models for predicting house prices with more precision. Linear regression, decision trees, and support vector machines are discussed as machine learning methods. Authors describe the mechanism through which these methods predict house prices. The significance of various variables, which are the geographical location of a property, its area, and the amenities available, is emphasized in the analysis of housing prices.

Chowhaan et al. (2023) analyzed the application of machine learning models for the prediction of house prices and compared Linear Regression, Lasso Regression, Support Vector Machines (SVM), Random Forest, and XGBoost. The study was able to demonstrate that Random Forest and XGBoost surpass standard regression

models since they are able to cope with sophisticated, non-linear housing data relationships. The size of the property, the number of bedrooms and bathrooms, and the location geographically all play a critical role in determining the prices of residential properties. Additionally, the proximity to city centres and public transport facilities matter. This work shows that ensemble learning methods can be used to improve the accuracy of price estimates. It also shows the worth of choosing the optimal features, preparing data correctly, and tuning model parameters to optimize the efficacy of machine learning models.

Fan et al. (2018), in their study, employed a range of machine learning techniques, which are Lasso and Ridge regression, Support Vector Regression (SVR), Random Forest, and Extreme Gradient Boosting (XGBoost), for the prediction of housing prices. RMSE was the primary measure of performance of the study, and following the logarithmic adjustment to the test data, the RMSE was 0.12019, showing hardly any overfitting and outstanding model performance. The area of a property, the parking lots available for a property, the bathrooms available for a property, and the age of a building are all significant factors when it comes to calculating the value of a property. These characteristics of a property significantly contribute to enhancing the prediction accuracy. They highlight the need to choose the right details when developing machine learning models to predict correct real estate prices.

Tekin and Sari (2022) applied machine learning to predict Istanbul real estate prices based on over 30,000 records of houses. They tried linear regression, polynomial regression, decision trees, random forests, and XGBoost and compared them based on R^2 and MAPE. The results showed that Random Forest and XGBoost outperformed other models, with the highest accuracy of MAPE approximately 20%, while polynomial regression gave unreliable results. Eliminating districts that had incomplete data also enhanced model performance. Their study demonstrates the importance of having good-quality data to make precise predictions.

2.2 Feature Importance Analysis

Truong et al. (2020) conducted a study on house price prediction using machine learning algorithms on the "Housing Price in Beijing" dataset, containing more than 300,000 data instances from the years 2009-2018. Random Forest, XGBoost, LightGBM, Hybrid Regression, and Stacked Generalization Regression were implemented with RMSLE as a model comparison evaluation approach. The research identified the most accurate model at predicting house prices for new data as Stacked Generalization Regression since it had the lowest RMSLE. The Random Forest model, while performing extremely well on the data it had been trained on, had a problem known as overfitting which it was not able to predict new data as accurately. Hybrid Regression, comprising algorithms such as Random Forest, XGBoost, and LightGBM, was the top-performing algorithm by ideally balancing bias and variance trade-off, making its predictions most reliable. Additionally, it was found that attributes such as the geographical location of a property, age, and size play a significant role in the market value of a property. The properties closer to city regions are more expensive. The study demonstrated the applicability of newer models like Stacked Generalization Regression and Hybrid Regression in predicting house prices, which can be effective tools for policymakers and investors. They can be applied to intelligent investment and decision-making in residential markets.

Gao et al. (2022) employed machine learning algorithms for predicting property prices in Greater Sydney, Australia, using real estate transactions and census data.

It benchmarked OLS, Ridge, Lasso, SVR, Decision Trees, Random Forest, Gradient Boosting, XGBoost, and Neural Networks and evaluated models based on R squared(R^2), MAPE, and MdAPE. It was found that Gradient Boosting and Random Forest were better than other models, and sub-area models (SA3 level) were more accurate than citywide models. Feature importance analysis identified number of bedrooms, floor area, and location features as primary drivers of price, and distances to city centers and transportation hubs as strongly significant. It explains how dividing regions into chunks and the use of mixed learning models help in providing more accurate estimates of real estate values. This is very helpful in city planning, how cities grow, and wiser investments in property.

Xu and Nguyen (2022) forecasted Chicago suburban prices using Redfin-collected data from 2018 to 2022. They tried linear regression, support vector regression, decision tree regression, random forest regression, and XGBoost regression, and they compared model performance using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). The results showed that XGBoost and random forest were best among other models, since they were able to capture complex relationships in the volatile post-pandemic market. Variable importance was taken into account using Shapley Values (SHAP), and it revealed square footage, yearly tax value, and number of bathrooms to be the most prominent variables for price prediction. This study illustrates that machine learning can be used in tracking market trends and helping real estate professionals and investors make informed decisions.

In Abdul Khani and Mohd Shafie (2023), various characteristics of a house are revealed to significantly contribute to housing prices. The researchers in this study identified that numerous factors associated with a house can influence its market price. These are the number of bathrooms, floor area size, location, type of property, and whether it is furnished. Some of these characteristics are more significant than others. Location is a big one because it decides how affordable a house is and how many people want to buy it in an area. The type of property and some of the building features can also cause prices to vary. This information helps buyers or investors in real estate make well-informed choices.

Zainal et al. (2020) investigated structural determinants of housing prices in Klang Valley, Malaysia, using SurveyMonkey questionnaires and logistic regression analysis. The findings indicated house prices were most influenced by bedrooms, bathrooms, and parking lots. Bigger houses with additional bedrooms and bathrooms are sought after since Malaysian families want a lot of living space. Car parking space is also crucial since most families have more than one car. However, the living room size is not as crucial since customers prioritize how space is maximized rather than having a big room. These are significant in emphasizing the need to have homes structured according to customers' wishes to effectively influence market prices.

2.3 Specialized Market and Property Type Applications

Lam Tatt et al. (2015) analyzed determinants of condominium prices near nascent MRT stations in Kuala Lumpur using hedonic regression analysis of 377 sale records between 2010 and 2013, provided by the Valuation and Property Services Department (JPPH). The study concluded that unit floor level and floor area were the structural determinants that had the largest impacts on prices. Larger floor areas strongly positively correlated with property price, since buyers preferred higher units, in accordance with Lin and Hwang (2003). Top-floor units were also asking top prices for the likely good view and fresh air. Higher-for-higher floors price increase is not unusual. In using multiple linear regression, the research ensured that these characteristics are factors which contribute to direction of change in prices. This establishes the importance of such characteristics in establishing property values and helping developers determine prices for condominiums in city centers.

Jamil et al. (2020) employed machine learning algorithms in forecasting green building costs in Kuala Lumpur, Malaysia, based on real estate data from the Green Building Index. The research employed various types of prediction models, including Linear Regression, Decision Tree, Random Forest, Ridge regression, and Lasso regression. The models were trained and tested on 80:20 train-test ratio, and their efficiencies were compared using R-squared statistics and Root Mean Squared Error (RMSE). Decision Tree Regressor performed the best followed by Random Forest, and Ridge and Lasso gave moderate efficiencies. Auto hyperparameter

tuning was also employed in this research with GridSearchCV for enhancing the efficiency of the model, and robust predictions of green building prices are achievable. Despite all the data that were gathered from these studies, there is still a research gap that has never been investigated.

2.4 Summary of Literature Review

Most existing studies on house price prediction employ filter-based or wrapper-based feature selection separately, which can lead to suboptimal results. Filter methods like IG rank by relevance and not redundancy, while wrapper methods like SVM-RFE are computationally expensive. This paper integrates both the approaches to retain the most relevant features and remove redundant features. Further, previous work scarcely examines model robustness against noisy and irrelevant features, assuming datasets are clean. This study mitigates this constraint by introducing noise-based, shuffled, and random categorical features to test the stability of the hybrid method in actual environments.

CHAPTER 3

METHODOLOGY

3.1 Information Gain

Information Gain is an important concept from information theory used in machine learning. It helps in picking the best features for tasks like classification and regression. Information Gain shows how much less uncertain we are about a target variable *Y* after looking at another variable *X*. In this study, Information Gain was applied to find the most important factors that affect house prices. By selecting these key features, the predictive accuracy of the machine learning models on house prices was improved significantly. This means the models can make better guesses about how much houses will cost.

$$IG(X,Y) = H(X) - H(X|Y) \quad (1)$$

Information Gain (Zhang et.al., 2024) is defined as the difference between the entropy of the target variable before and after the observation of the feature which is shown in equation 1, where $H(X)$ is the entropy of the target variable *X*, representing the uncertainty inherent in predicting *X* without any additional information, and $H(X|Y)$ is the conditional entropy of *X* given *Y*, representing the remaining uncertainty after knowing *Y*.

$$H(X) = -\sum_{x_i \in X} P(x_i) \log\big(P(x_i)\big) \tag{2}$$

Entropy $H(X)$ is calculated using equation 2 where $P(x_i)$ is the probability of occurrence of outcome x_i out of all possible outcomes n. Entropy quantifies the expected value of the information contained in the data.

$$H(X|Y) = -\sum_{y_j \in Y} P(y_j) \sum_{x_i \in X} P(x_i|y_j) \log\Big(P\big(x_i|y_j\big)\Big) \tag{3}$$

Conditional entropy is calculated using equation 3 where $P(y_j)$ represents the probability of different values of Y and $P(x_i|y_j)$ represents the conditional probability of X given Y.

Mutual information, a generalization of information gain, is utilized in regression studies involving continuous variables, including forecasting house prices. Without supposing any distribution, mutual information captures both linear and non-linear interactions by quantifying the amount of information one variable possesses about another (Ross, 2014; Vergara & Estevez, 2014). It is particularly useful in feature

selection because it can detect any kind of dependency between variables, not just linear correlations.

$$MI(X;Y) = H(X) - H(X|Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (4)$$

Mutual information between variables *X* and *Y* is shown in equation 4 where $p(x,y)$ is the joint probability distribution function of *X* and *Y*, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of *X* and *Y* respectively (Alalhareth & Hong, 2023).

IG has limitations, particularly in handling feature redundancy. Since IG evaluates each feature independently without considering interdependence, it may select multiple features that convey similar information, leading to an overly complex model (Brown et al., 2012).

3.2 SVM-RFE

Support Vector Machine Recursive Feature Elimination, which is simply SVM-RFE, is an important feature selection method applied in data analysis to determine significant features of data needed to perform activities such as classification and regression. An example of this is using it to predict the prices of houses. The process is a method of eliminating the least important features by looking at the weights assigned to every feature by SVM. These weights measure a feature's contribution towards making predictions (Lin et al., 2012). By only considering the most important features, SVM-RFE minimizes the dataset. It makes predictions more accurate and decreases the model's complexity, preventing overfitting. It also simplifies the model. SVM-RFE is most useful in real estate and other fields in which data is always complicated with a multitude of variables. It also works fine when dealing with data sets containing a huge number of features but few observations. Steps involved in the process are:

1. Train the SVM Model: The whole dataset is trained using an SVM model, and feature weights are calculated.

2. Rank Features: Features are ranked in terms of the absolute values of their weights. Features with lower weights have lower weighted values.

3. Remove Features: The feature (or a group of features) with the least significance is removed from the dataset.

4. Repeat: Perform the same operation on the filtered dataset until we reach our desired number of features.

This recursive process ensures that the most informative features are retained while the redundant features are eliminated, leading to a more effective and efficient predictive model.

SVM-RFE has some limitations, primarily its high computational complexity (Guyon et al., 2002). The computational expense of Support Vector Machine Recursive Feature Elimination comes from the necessity of multiple iterations of training an SVM model in every process of the feature removal process, hence making it a resource-intensive technique. It becomes especially demanding where large data or complex real estate data containing many details come into play (Sanz et al., 2018). Therefore, SVM-RFE can be a time-consuming and processing-demanding process. This is a serious limitation regarding real-time usage when speedy outcomes are necessary. In an effort to speed things up, it often needs powerful computers or efficient ways of spreading the work out to many computers. This is a restriction as to how and where SVM-RFE can be utilized, particularly in situations where one wants answers straight away.

3.3 Hybrid Method

This study employs a hybrid feature selection approach, combining Information Gain with Support Vector Machine Recursive Feature Elimination to overcome both limitations. This approach enhances model performance by leveraging the strengths of both filter and wrapper methods. Prasetiyowati et al. (2021) mention that IG can efficiently achieve lower prediction errors and higher accuracy by removing irrelevant features, while SVM-RFE further refines the selection by eliminating redundant and less impactful features. The combination ensures that the final feature subset maximizes predictive power while maintaining computational efficiency. The goal was to identify the most relevant features while reducing redundancy and computational costs. Below is the flow of the methodology:

Steps:

1. Load and preprocess the dataset to ensure data consistency.
2. Calculate the mutual information (MI) score. This provides a score indicating the amount of information each feature shares with the target. Features with higher scores were considered more informative.
3. The features were ranked based on their mutual information scores in descending order. Higher scores indicated a stronger association with the target variable. (Chen & Guestrin, 2016).
4. The top 15 features were selected based on MI scores, balancing model performance, and complexity. The selection process was repeated 100 times, retaining only features that appeared in at least 95 times to ensure stability.
5. SVM-RFE was applied to the selected features to further refine the selection.

6. The SVM-RFE process was repeated multiple times to ensure the stability of selected features. Features that consistently ranked high across iterations were retained for model training.
7. Several regression models, including Random Forest, Gradient Boosting, Support Vector Regression (SVR), Ridge, Lasso, and Linear Regression, were trained using the final feature set.
8. Model performance was evaluated using 5-fold cross-validation with metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). Repeat the entire methodology five times, averaging the results of all metrics across these repetitions to ensure robustness and reliability. The model with the lowest RMSE was selected as the best-performing model.

Various models were used in this study to assess how well they predicted house prices. A baseline model was first developed without applying any feature selection to serve as a reference point. Then, three models were tested to compare their impact on model performance. The initial model employed only Information Gain to examine and choose the most appropriate features according to their correlation with house prices. Although Information Gain performed well, it did not take into account feature dependence, which could result in the selection of redundant features. The subsequent model employed a hybrid feature selection strategy that integrated Support Vector Machine Recursive Feature Elimination and Information Gain to overcome this drawback. The third model applied the same hybrid

technique to an altered dataset with incorporated noise to ascertain the stability of the hybrid strategy. The new data set was achieved by appending five numerical features of Gaussian noise, three permutations of original features, and two randomly created categorical features. This was carried out to identify the capability of the hybrid approach in dealing with irrelevant features and model degradation defense.

3.4 Hyperparameter Tuning and Performance Metrics

Several models were developed and compared in the current study with the aim of making house price predictions based on some attributes. Models used include the Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regression (SVR), Ridge Regression, Lasso Regression, and Linear Regression. Each model's performance was experimented with using a 5-fold cross-validation technique, a method reducing the likelihood of overfitting by splitting the training data into subsets and the model into various divisions of the data. The performance measure and the measure selected for analysis were Mean Squared Error (MSE) as the performance measure and Root Mean Squared Error (RMSE) because it is more interpretable with better acceptability since RMSE has general acceptability as a default in regression analysis to measure accuracy of prediction (Chai & Draxler, 2014). The score with the least RMSE was selected as the best score and further chosen for hyperparameter tuning to improve its predictive power (Varma & Simon, 2006).

To improve the performance of the model, hyperparameter tuning was performed using RandomizedSearchCV because it performs better than exhaustive grid search, particularly for large datasets and complicated models (Bergstra & Bengio, 2012). Hyperparameter selection was undertaken using values commonly used in journal publications. For Random Forest Regressor, hyperparameter tuning was performed using the number of estimators, maximum depth, minimum samples split, and minimum samples leaf (Breiman, 2001). Gradient Boosting Regressor was tuned

by changing the number of estimators, learning rate, and maximum depth (Friedman, 2001). In SVR, significant hyperparameters were the regularization parameter C, epsilon in the loss function, and kernel type (Smola & Schölkopf, 2004). By classically testing various regression models and optimizing the best regressor based on performance metrics, this study ensured that the derived model achieved increased accuracy and generalizability to new data.

CHAPTER 4

RESULT AND DISCUSSION

4.1 Dataset Introduction and Data Preprocessing

The dataset used in this study was collected from Mudah.my, a popular Malaysian online marketplace, and consists of house listings with various attributes related to property characteristics and pricing. The dataset can be accessed and downloaded from https://www.kaggle.com/datasets/mcpenguin/raw-malaysian-housing-prices-data. This study looks at how different features affect house prices using a data set with 32 features. These features cover both numbers and categories. The quantitative characteristics include things like the size of the house in square meters, and the number of bedrooms and bathrooms. The qualitative characteristics include details like the type of house, where it is situated, and what materials it is constructed of. Features related to amenities, accessibility, and the condition of the house are others. All these facts inform us about house prices. Data cleaning was done to deal with missing values, delete duplicates, and make categorical variables consistent. Feature encoding methods, like label encoding, were used on categorical variables to enable them to be used in machine learning models. Label encoding was used over one-hot encoding for two primary reasons. First, label encoding is quicker and simpler as it translates category data into numbers immediately. This is helpful where there are numerous different categories. One-hot encoding would add many more columns, which would bloat the data and processing and be costly. Second, label encoding does not destroy the natural order of categories, while one-

hot encoding treats all categories equally, which could interfere with their natural order (Seveso et al., 2020).

The dataset was randomly divided into two subsets in a 7:3 ratio, representing the training and testing sets, respectively. The sample consists of a complete feature set $X = \{x_1, x_2, \dots, x_N\}$ where N represents the total number of selected features. Before model training, the dataset underwent normalization to scale numerical features within the range [0,1] to ensure consistency and improve model performance. Min-max normalization was chosen over z-score normalization because it preserves the original distribution of the data (Saranya, 2013). Additionally, it is beneficial when numerical features have varying ranges, preventing any one feature from dominating the learning process. The dataset underwent extensive preprocessing to enhance its quality and ensure compatibility with machine learning models. Several unnecessary columns, such as developer and firm-related attributes, were removed to retain only relevant features. Missing values in continuous variables were imputed using the mean strategy, while ordinal and nominal variables were imputed with the most frequent values. Mean imputation was chosen for continuous variables due to its computational efficiency and suitability for normally distributed data, as it preserves the dataset's overall structure (Baraldi & Enders, 2010). The recreational facilities feature was categorized into a binary format, indicating the presence of amenities such as swimming pools, playgrounds, and gyms. Similarly, security and access features, including gated security and lift accessibility, were encoded to distinguish properties with enhanced security. The state and area were

extracted from the address field to provide structured location-based features, aiding in regional price analysis. The age of the building was computed by subtracting the completion year from 2024, transforming it into a numerical variable for better interpretability. Additionally, binary encoding was applied to features like proximity to schools, malls, hospitals, highways, bus stops, parks, and railway stations to reflect accessibility factors influencing house prices.

4.2 Results and Discussion

Table 4.2.1: The selected features by information gain

No	Model 1,2 & 3
1	# of Floors
2	Age of building
3	Area
4	Bathroom
5	Bedroom
6	description
7	Parking Lot
8	Property Size
9	Property Type
10	Recreational Facilities
11	Security and Access
12	State
13	Total Units

Table 4.2.1 shows the selected features using Information Gain (IG) for the three scenarios. The selected features remain consistent across all scenarios, indicating that IG identifies the same highly relevant features regardless of additional noise or different feature selection approaches. These features include the number of Floors,

Age of Building, Area, Bathroom, Bedroom, Property Size, Property Type, Recreational Facilities, Security and Access, State, and Total Units, among others.

Table 4.2.2: Feature Selected using SVM-RFE

No	Model 2 &3
1	# of Floors
2	Age of building
3	Bathroom
4	Bedroom
5	Parking Lot
6	Property Size
7	Property Type
8	Recreational Facilities
9	Security and Access
10	State
11	Total Units

After applying SVM-RFE to the selected features from Information Gain, some features were removed in both models 2 and 3, which are shown in Table 4.2.2 to refine the selection. In models 2 and 3, two features, which are "Description" and "Area" were eliminated, indicating that these features contributed less to the predictive performance of the model. The removal of these features highlights how SVM-RFE focuses on eliminating redundant or less influential variables, ensuring

that only the most relevant attributes remain for the model to make accurate predictions. By combining both methods, the hybrid approach enhances feature selection by ensuring that highly relevant features are included while simultaneously eliminating redundant or less impactful features.

Table 4.2.3: Performance metrics result of regressors in baseline (no feature selection)

Regressors	MSE	RMSE	R^2 Squared
Random Forest	**3.36E+10**	**1.81E+05**	**0.6684**
Gradient Boosting	**3.34E+10**	**1.79E+05**	**0.6789**
SVR	1.07E+11	3.23E+05	-0.0446
Ridge Regression	2.51E+11	3.84E+05	-0.8380
Lasso	2.52E+11	3.84E+05	-0.8390
Linear Regression	2.52E+11	3.84E+05	-0.8390

In the baseline scenario with no feature selection, as shown in Table 4.2.3, the performance of six regression models, which are Random Forest, Gradient Boosting, SVR, Ridge Regression, Lasso, and Linear Regression, varied significantly. Random Forest and Gradient Boosting stood out with strong predictive capabilities, achieving R-squared values of 0.6684 and 0.6789, respectively, alongside relatively low Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). These results show that ensemble techniques are good at handling datasets with irrelevant or redundant features and can detect underlying

patterns. SVR, Ridge Regression, Lasso, and Linear Regression, however, performed poorly, with negative R^2 scores and large error metrics, which are indicative of their failure to generalize when handling unfiltered data. This discrepancy shows the vulnerability of linear models and SVR to irrelevant features, which introduce noise and reduce accuracy. The findings emphasize the value of feature selection as a crucial step to improve model reliability, especially for weaker performers.

Table 4.2.4: Comparison of results between feature selection models

Regressors	Performance Metrics	Feature Selection		
		Model 1 (Information Gain)	**Model 2 (Hybrid Method on Original Dataset)**	**Model 3 (Hybrid Method on noise dataset)**
Random Forest	RMSE	**185518.5223**	**154403.7**	**154403.7**
Gradient Boosting		187999.0826	169232.6	169232.6
SVR		323025.9463	355854	355854
Ridge Regression		385640.1226	248454	248454
Lasso		385716.666	248438.8	248438.8
Linear Regression		385717.9885	248438.4	248438.4
Random Forest	R-squared	**0.6522**	**0.8008**	**0.8008**
Gradient Boosting		0.6464	0.7608	0.7608
SVR		-0.0445	-0.0577	-0.0577
Ridge Regression		-0.8527	0.4844	0.4844
Lasso		-0.8537	0.4845	0.4845
Linear Regression		-0.8537	0.4845	0.4845

Table 4.2.4 shows the comparison results among the feature selection models. In model 1, where only Information Gain was used for feature selection, the overall model performance declined compared to model 2. Random Forest is the best-performing regressor, followed closely by Gradient Boosting, though both exhibited slightly reduced accuracy. The negative R-squared values for SVR, Ridge Regression, Lasso, and Linear Regression indicate that these models failed to explain variance in the target variable and performed worse than a simple mean prediction. The results suggest that relying solely on Information Gain for feature selection is less effective than the hybrid approach, as it may not sufficiently eliminate redundant or irrelevant features. This reinforces the advantage of combining Information Gain with SVM-RFE, which is model 2, enhancing model robustness and predictive accuracy.

Model 2 results evidently indicate that hybrid feature selection process (Information Gain + SVM-RFE) improves model performance in a satisfactory way compared to the baseline where no feature selection was used. Random Forest still maintained the best regression performance, showing robust predictive accuracy with minimal overfitting levels. Good performances were generated by Gradient Boosting too but with predictively higher error rates. Compared to the baseline, Random Forest and Gradient Boosting both outperformed it in terms of prediction accuracy, indicating that removal of irrelevant features enhances the performance of models. Support Vector Regression (SVR) still underperformed in the sense that it could not identify significant patterns in the data. Ridge Regression, Lasso, and

Linear Regression were performing decently but the regressors still underperformed in identifying complex patterns.

The results of model 3 confirm that even with the incorporation of noise features, model performance in general was still strong, particularly for Random Forest and Gradient Boosting. This indicates that the hybrid feature selection method efficiently retained the most important features, thus making the models more noise resistant. Random Forest came out top once more, with minimal overfitting and perfect predictive accuracy, then Gradient Boosting, which was similarly well-performing but with slightly more error. SVR couldn't learn meaningful relationships, but the consistency with which the tree-based approaches came in high points shows that the hybrid approach had successfully removed unnecessary features without compromising necessary ones. This shows the efficacy of the hybrid feature selection technique in being able to make good predictions, even when the data contains noise.

Feature selection is an essential procedure in machine learning in in-house price prediction scenarios since it improves model performance through the determination of the best predictors. The results have proved the great advantage of employing SVM-RFE, particularly its ability to improve model performance. SVM-RFE application, through selective elimination of less informative features, enhances accuracy, generalizability, and interpretability (Sanz et al., 2018). With

this process, models can attain superior performance measures, including decreased prediction errors and increased R-squared values. SVM-RFE also reduces the risk of overfitting since redundant features are eliminated that could bring about noise, hence enabling the model to generalize suitably to forthcoming data (Li et al., 2024). In addition, the picked variables are simpler to interpret, and real estate stakeholders can easily comprehend key drivers of price.

The result of the research indicates that the implementation of a hybrid technique using Information Gain (IG) and Support Vector Machine Recursive Feature Elimination (SVM-RFE) proves to be superior to the use of Information Gain. Specifically, Information Gain is used for important feature identification, while SVM-RFE fills in the gap in eliminating redundant features or less prominent features. This combination enhances the predictive ability of the model. Even though introducing additional noise into the data, the hybrid model performed well, hence bearing witness to the robustness of the model and the capability to ignore noisy or irrelevant information. This therefore bears witness to the hybrid approach as being a robust feature selection method.

The attributes chosen by the hybrid attribute selection technique, being the fusion of Information Gain (IG) and Support Vector Machine Recursive Feature Elimination, are number of parking lots, number of bathrooms and bedrooms, number of floors, property size, property type, recreational facilities, security and

access, state, and total units. These characteristics were consistently identified as the most relevant in different datasets, such that the predictive model focuses on significant factors influencing house prices and eliminates redundant or less important attributes. This specific selection improves the model to be more precise and specific by keeping only the most useful information.

Table 4.2.5 Result of Hyperparameter tuned on Hybrid Method

	Before Hyperparameter tuned	After Hyperparameter tuned
RMSE	154403.7	153977.4
R-squared	0.8008	0.802

The model's performance improved after applying hyperparameter tuning to the Random Forest model using the hybrid feature selection method. The Root Mean Squared Error (RMSE) decreased from 154403.7 to 153977.4, while the R-squared value increased from 0.8008 to 0.802. These improvements indicate that fine-tuning the model’s parameters enhanced its predictive accuracy, though the changes were marginal.

The optimal hyperparameters selected for the Random Forest model included n_estimators = 300, min_samples_split = 2, and min_samples_leaf = 1. Increasing n_estimators to 300 added robustness to the model by averaging predictions over a larger ensemble of trees, reducing variance and output stabilization. Probst et al.

(2019) found that higher values of n_estimators are about to increase performance, though improvements decrease after a certain point. Similarly, Oshiro et al. (2012) suggested that model performance plateaus from 128 to 256 trees, with further increases having diminishing returns. Here, increasing n_estimators to 300 likely provided some marginal improvement beyond this point, again stabilizing the predictions.

By setting min_samples_split to 2, its lowest possible value, deeper tree growth was allowed, and the model could capture more intricate patterns within the data. Although smaller values do increase the danger of overfitting, Random Forest's ensemble method prevented this by averaging over 300 trees, which generalized to new data. Additionally, min_samples_leaf = 1 permitted single-sample leaf nodes, reducing bias by allowing highly detailed tree structures. Segal (2004) emphasized that such low values improve model flexibility, which proved beneficial given the dataset’s complexity and the effectiveness of the hybrid feature selection method in reducing irrelevant features.

CHAPTER 5

CONCLUSION

5.1 Summary of Research

Within the research, the prediction of house price is enhanced based on the hybrid feature selection technique created for the strategy, which entails information gain to position features according to how well they may suit the task at hand and Support Vector Machine Recursive Feature Elimination, a recursive feature elimination in which unnecessary and least vital features are progressively eliminated. The combination of these two techniques was to look for the best set of features that would improve how good the prediction model works. It was successful in that it made a model which could predict with accuracy. Another very important finding was that the model remained robust in noise dataset. Therefore, this showed the strength of the model and how it can be ignored if some information is unnecessary.

The hybrid feature selection approach has outperformed the information gain alone, thus establishing the strength of having an integrated multi-feature selection strategy. The hybrid produced a better R-squared value from 0.6522 to 0.8008, which represents a 22.8% increase in predictive accuracy. The Root Mean Squared Error (RMSE) decreased from 185,518.52 to 154,403.7, meaning that there was an actual reduction of 16.77% from the error of prediction. This improved

performance was because it had the ability to remove redundancy and also rank features, resulting in a drastically improved, more precise model. From these findings, it is essential that a combination of multiple methodologies is necessary to achieve the most precise, competent results for using predictive modeling.

It is quite helpful in research for people dealing in real estate. With the improved model, investors can spot properties, which they feel are being sold at prices less than their appraisal values. This allows the individuals to make more informed decisions that minimize risks in their property investments and manage property portfolios more effectively. This information could also be used by policymakers and regulators as they develop taxes that are tax-fairly imposed while anticipating changes that can happen within the housing market. Homebuyers also have an advantage where they receive far more precise appraisals, therefore not getting overcharged or undercharged for a home. The findings are crucial in real estate industries, as accurate price predictions would enable real decision-making within this domain. Overall, the combination of various techniques in this study enhances the performance and reliability of the model in real-life scenarios.

5.2 Limitation of Hybrid Method

When used on time-dependent or streaming data, the hybrid feature selection method of Information Gain and SVM-RFE shows notable drawbacks. SVM-RFE does not take into consideration changing patterns over time because it is naturally built for static datasets. Because of this, it is inappropriate for dynamic settings like the real estate market, where information like listings and pricing are subject to regular changes. The chosen characteristics could soon become out of date without frequent upgrades, which would impair the model's functionality. Furthermore, SVM-RFE's effectiveness in predicting future market trends or adjusting to changing housing price dynamics on platforms such as Mudah.my is limited due to its inability to capture temporal dependencies or sequential linkages.

5.3 Future Work

Follow-up work on the application of a hybrid feature selection approach to residential property value prediction can be undertaken in several interesting directions. Firstly, it is important to conduct cross-validation of the model's efficacy with different datasets reflecting real-world variations. This entails experimentation on datasets from different places with different economic profiles and with differing levels of information on properties. Working in such a direction will offer insight into the generalizability of the methodology to different property markets.

Also, feature selection can be optimized by applying advanced techniques such as genetic algorithms or principal component analysis (PCA). These newer techniques can aid in improving speed and efficiency. Management of the complicated calculations of the hybrid method is also crucial. This observation is most applicable to the SVM-RFE module, whose optimization is achievable through parallel processing or approximating techniques to deal with large datasets. Also significant is the fact that the model's predictions need to be explainable. With explainable artificial intelligence methods such as SHAP values or LIME, one can visualize the top determinants of house prices. It would enhance understanding and gain trust in the predictions for everyone involved.

Hyperparameter optimization can be improved by employing techniques like Bayesian optimization or genetic algorithms. The above methods can automate

procedures, enable the determination of best parameter values easily, and are cost-saving. In the future, there is a possibility of converting the model to an online application with real-time operation. This would allow real estate agents to enter the specifications of a property and receive quick, data-driven price appraisals. With these improvements, the study can play a part in creating more accurate and transparent models for the prediction of housing prices. This innovation would allow for better decision-making in the areas of policy making and real estate investment.

REFERENCES

Abdul Khani, M.A. and Mohd Shafie, S.A. (2023). 'Factors affecting the house price among Kuala Lumpur, Selangor and Johor', in *Proceedings of the International Jasin Multimedia & Computer Science Invention and Innovation Exhibition (i-JaMCSIIX 2023)*. Universiti Teknologi MARA, pp. 102–105. Available at: https://ir.uitm.edu.my/id/eprint/94364/ [Accessed: 30 Jun. 2024].

Abdul-Rahman, S., Mutalib, S., Zulkifley, N.H. and Ibrahim, I., 2021. Advanced machine learning algorithms for house price prediction: Case study in Kuala Lumpur. *International Journal of Advanced Computer Science and Applications*, 12(12).

Alalhareth, M. and Hong, S.C., 2023. An improved mutual information feature selection technique for intrusion detection systems in the Internet of Medical Things', *Sensors*, 23(4971). Baraldi, A.N. and Enders, C.K. (2010). 'An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), pp. 5–37.

Bergstra, J. and Bengio, Y., 2012. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1), pp. 281–305.

Breiman, L., 2001. Random forests. *Machine Learning*, 45, pp. 5–32.

Brown, G., Pocock, A., Zhao, M.J. and Luján, M., 2012. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13(1), pp. 27–66.

Chai, T. and Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7, pp. 1247–1250.

Chen, T. and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, pp. 785–794.

Chowhaan, M. et al., 2023. Machine learning approach for house price prediction. *Asian Journal of Research in Computer Science*, 16(2), pp. 54–61.

Fan, C., Cui, Z. and Zhong, X., 2018. House prices prediction with machine learning algorithms. *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, pp. 6–10.

Friedman, J.H., 2000. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), pp. 1189–1232.

Furia, P. and Khandare, A., 2022. Real estate price prediction using machine learning algorithms. *Advanced Analytics and Deep Learning Models*. Wiley.

Gao, Q., Shi, V., Pettit, C. and Han, H., 2022. Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia. *Land Use Policy*, 123, p. 106409.

Guo, J., 2023. Feature selection in house price prediction. *Highlights in Business, Economics and Management*, 21, pp. 746–752.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46, pp.389-422.

Ho, W.K., Tang, B.-S. and Wong, S.W., 2020. Predicting property prices with machine learning algorithms. *Journal of Property Research*, 37(4), pp. 297–320.

Jamil, S., Mohd, T., Masrom, S. and Ab Rahim, N., 2020. Machine learning price prediction on green building prices. *2020 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, pp. 1–6.

Kostic, Z. and Jevremovic, A., 2021. What image features boost housing market predictions? *arXiv preprint arXiv:2107.07148*.

Lam Tatt, S., Ng Po Yi, M. and Fuey Lin, A., 2015. Factors affecting prices of condominiums nearby developing MRT stations in Kuala Lumpur. *21st Annual Pacific-Rim Real Estate Society Conference*, Kuala Lumpur, Malaysia.

Li, J., Wang, J. and Xue, E., 2024. Applying a support vector machine (SVM-RFE) learning approach to investigate students' scientific literacy development: Evidence from Asia, Europe, and South America. *Journal of Intelligence*, 12(111).

Liang, J. and Yuan, C., 2021. Data price determinants based on a hedonic pricing model,' *Big Data Research*, 25, p. 100249.

Lin, X. et al., 2012. A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *Journal of Chromatography B*, 910, pp. 149–155.

Mathotaarachchi, K.V., Hasan, R. and Mahmood, S., 2024. Advanced machine learning techniques for predictive modeling of property prices. *Information*, 15(6), p. 295.

Oshiro, T.M., Perez, P.S. and Baranauskas, J.A., 2012. How many trees in a random forest?. *Machine Learning and Data Mining in Pattern Recognition*, 7376, pp. 154–168.

Prasetiyowati, M.I., Maulidevi, N.U. and Surendro, K., 2021. Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. *Journal of Big Data*, 8(84).

Probst, P., Wright, M.N. and Boulesteix, A.-L., 2019. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), p. e1301.

Ross, B.C., 2014. Mutual information between discrete and continuous data sets. *PLoS ONE*, 9(2), p. e87357.

Sanz, H., Valim, C., Vegas, E., Oller, J.M. and Reverter, F., 2018. SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics*, 19(432).

Saranya, C. and Santhamoorthy, E., 2013. A Study on Normalization Techniques for Privacy Preserving Data Mining. *International Journal of Engineering and Technology (IJET)*, 5(3), pp. 2701–2704.

Segal, M.R., 2004. Machine learning benchmarks and random forest regression. *Center for Bioinformatics & Molecular Biostatistics*, 5(1), pp. 1–15.

Seveso, A., Campagner, A., Ciucci, D. and Cabitza, F., 2020. Ordinal labels in machine learning: a user-centered approach to improve data validity in medical settings. *BMC Medical Informatics and Decision Making*, 20(S5).

Shi, S., Mangioni, V., Ge, X. J., Herath, S., Rabhi, F. and Ouysse, R., 2021. House price forecasting from investment perspectives. *Land*, 10(10), p. 1009.

Sharma, S., Kumari, S., Goyal, S. and Nirala, R., 2024. A review: Real estate price prediction using machine learning with research and trends. *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, pp. 1239–1244.

Smola, A.J. and Schölkopf, B., 2004. A tutorial on support vector regression. *Statistics and Computing*, 14, pp. 199–222.

Tekin, M. and Sari, I.U., 2022. Real estate market price prediction model of Istanbul. *Real Estate Management and Valuation*, 30(4), pp. 1–16.

Truong, Q., Nguyen, M., Dang, H. and Mei, B., 2020. Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 174, pp. 433–442.

Vargas-Calderón, V. and Camargo, J. E., 2022. Towards robust and speculation-reduction real estate pricing models based on a data-driven strategy. *Journal of the Operational Research Society*, 73(12), pp. 2794–2807.

Varma, S. and Simon, R., 2006. Bias in error estimation when using cross-validation for model selection', *BMC Bioinformatics*, 7(91).

Vergara, J.R. and Estevez, P., 2014. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 25, pp. 123–141.

Wang, F., Zou, Y., Zhang, H. and Shi, H., 2019. House price prediction approach based on deep learning and ARIMA model. *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, pp. 303–307.

Wang, H. and Xu, Y., 2019. Variable selection in high-dimensional linear regression: Persistence of the forward stepwise selection. *Journal of Applied Statistics*, 46(1), pp. 1–18.

Wei, C. *et al.*, 2022. The research development of Hedonic Price Model-Based Real Estate Appraisal in the era of Big Data. *Land*, 11(3), p. 334.
Xu, K. and Nguyen, H., 2022. Predicting housing prices and analyzing real estate markets in the Chicago suburbs using machine learning. *Journal of Student Research*, 11(3).

Yazdani, M., 2021. Machine learning, deep learning, and hedonic methods for real estate price prediction. *arXiv preprint arXiv:2110.07151*.

Yee, L.W. et al., 2021. Using machine learning to forecast residential property prices in overcoming the property overhang issue. *2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pp. 1–6.

Zainal, R., Mat Radzuan, I.S. and Hassan, N.H., 2020. Factors influencing home buyers' purchase decisions in Klang Valley, Malaysia. *Malaysian Journal of Sustainable Environment*, 7(2).

Zhang, B. et al., 2024. Information gain-based multi-objective evolutionary algorithm for feature selection. *Information Sciences*, 677, p. 120901.

Appendices

Appendix A

Universiti Tunku Abdul Rahman			
Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1 of 1

FACULTY OF SCIENCE

Full Name(s) of Candidate(s)	Low Jun Liang
ID Number(s)	22ADB02353
Programme / Course	SCOR
Title of Final Year Project	Enhancing house price prediction by hybrid method: A combination of Information Gain and Support Vector Machine Recursive Feature Elimination

Similarity	**Supervisor's Comments** **(Compulsory if parameters of originality exceeds the limits approved by UTAR)**
Overall similarity index: 10 % **Similarity by source** Internet Sources: 8 % Publications: 5 % Student Papers: 0 %	
Number of individual sources listed of more than 3% similarity: 0	

Parameters of originality required and limits approved by UTAR are as follows:
- **(i) Overall similarity index is 20% and below , and**
- **(ii) Matching of individual sources listed must be less than 3% each , and**
- **(iii) Matching texts in continuous block must not exceed 8 words**

Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

Signature of Supervisor

Name: Dr Chin Fung Yuen

Date: 29 Apr 2025

FYP for turnitin.docx

ORIGINALITY REPORT

10%	8%	5%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International Conference on Algorithms in Advanced Artificial Intelligence (ICAAAI-2024)", CRC Press, 2025 Publication	<1%
3	www.frontiersin.org Internet Source	<1%
4	www.uyik.org Internet Source	<1%
5	rcsoccersim.github.io Internet Source	<1%
6	internationalpubls.com Internet Source	<1%
7	www.biorxiv.org Internet Source	<1%
8	www.nature.com Internet Source	<1%
9	arxiv.org Internet Source	<1%
10	link.springer.com Internet Source	<1%

11	Saiyed Salim Sayeed, Hemant Kumar Sharma, Pramod Kumar Yadav, Brijesh Mishra. "Advances in Electronics, Computer, Physical and Chemical Sciences - Proceedings of International Conference on Electronics, Computer, Physical and Chemical Sciences (ICECPCS 2024), July 19–21, 2024, JNRM Port Blair, India", CRC Press, 2025 Publication	<1 %
12	dspace.cvut.cz Internet Source	<1 %
13	ijsrem.com Internet Source	<1 %
14	Gregor Pavlin. "A probabilistic approach to resource allocation in distributed fusion systems", Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems - AAMAS 05 AAMAS 05, 2005 Publication	<1 %
15	Sai Kiran Oruganti, Dimitrios A Karras, Srinesh Singh Thakur, Janapati Krishna Chaithanya, Sukanya Metta, Amit Lathigara. "Digital Transformation and Sustainability of Business", CRC Press, 2025 Publication	<1 %
16	dallascrgt77543.pages10.com Internet Source	<1 %
17	export.arxiv.org Internet Source	<1 %
18	stce.huce.edu.vn Internet Source	<1 %
19	fas.vau.ac.lk Internet Source	<1 %

20	www.irjet.net Internet Source	<1 %
21	acc-ern.tul.cz Internet Source	<1 %
22	econpapers.repec.org Internet Source	<1 %
23	www2.mdpi.com Internet Source	<1 %
24	Asif Rahman, Faisal Bin Abdur Rahman, Anharul Islam, Ifrat Jahan, K.M.A. Salam. "Cerebral Stroke Prediction Using Machine Learning Algorithms", 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), 2023 Publication	<1 %
25	coek.info Internet Source	<1 %
26	digital.library.adelaide.edu.au Internet Source	<1 %
27	etheses.whiterose.ac.uk Internet Source	<1 %
28	github.com Internet Source	<1 %
29	www.journalssystem.com Internet Source	<1 %
30	catalog.okbu.edu Internet Source	<1 %
31	ieeexplore.ieee.org Internet Source	<1 %
32	repository.up.ac.za Internet Source	<1 %

33	www.researchgate.net Internet Source	<1 %
34	www.researchsquare.com Internet Source	<1 %
35	www.slideshare.net Internet Source	<1 %
36	A. Koike. "Comparison of methods for chemical-compound affinity prediction", SAR and QSAR in Environmental Research, 2006 Publication	<1 %
37	Pooja Jha, Shalini Mahato, Prasanta K. Jana, Sudhanshu Maurya, Inès Chihi. "Artificial Intelligence based Solutions for Industrial Applications", CRC Press, 2024 Publication	<1 %
38	file.techscience.com Internet Source	<1 %
39	uac.incd.ro Internet Source	<1 %
40	www.ncbi.nlm.nih.gov Internet Source	<1 %
41	B.K. Lavine. "Feature Selection: Introduction", Elsevier BV, 2009 Publication	<1 %
42	Iacopo Carnacina, Mawada Abdellatif, Manolia Andredaki, James Cooper, Darren Lumbroso, Virginia Ruiz-Villanueva. "River Flow 2024", CRC Press, 2025 Publication	<1 %
43	Victor Hugo Peres Silva, Carolina Luiza Emereciana Pessoa, Derica dos Santos Sousa, Ricardo Stefani. "Using synthetic data to develop machine learning models to predict	<1 %

	the performance of fiber- reinforced concrete", Springer Science and Business Media LLC, 2024 Publication	
44	Yifan Jiang, Disen Liao, Qiyun Zhu, Yang Young Lu. "PhyloMix: Enhancing microbiome-trait association prediction through phylogeny-mixing augmentation", Cold Spring Harbor Laboratory, 2024 Publication	<1 %
45	mdpi-res.com Internet Source	<1 %
46	mobt3ath.com Internet Source	<1 %
47	pdfs.semanticscholar.org Internet Source	<1 %
48	unaab.edu.ng Internet Source	<1 %
49	www.fi.muni.cz Internet Source	<1 %