1 Introduction

In this technological era, technologies are progressing at an incredible pace. Information is being produced at incredible volume. Many issues occurring in the world may be solved using big data techniques like machine learning techniques. Numerous researches conducted throughout the world shows that machine learning is a powerful tool that is capable of solving issues. Be it either in the side of prediction, classification, diagnosis, detection, or even recognition. Unfortunately, the usage of machine learning in house prices prediction in Malaysia is very sparse as compared to other country states of the world.

It is critical for a person who is either a buyer or seller in the house market to know the variables that correspond to the house pricing. It is even more important for such a person to be able to consider whether the price of the house is realistic to the aspects that it possesses.

Machine learning can solve the housing prices problem because of its ability to accurately predict from data. It is imperative to utilize machine learning properly to determine the significant dependents. As the performance metrics of the results produced are very dependent on the collected dataset and preprocessed data.

For this research, the chosen machine learning techniques are multiple linear, Bayesian ridge, decision tree and random forest. The chosen machine learning techniques are compared to get the performance scores in predicting house price as compared to its actual price.

2 Related Work

It is important to lay down several groundwork before fully committing to the project. It is essential to conduct a literature review. Related papers should be analyzed in regards to price prediction in house prices. Each paper will depict its technicalities and will be compared.

Dagar and Kapoor [1] study on house price prediction based on the features selected. From their findings, they concluded that regression analysis is the best solution to the problem that the authors are addressing. The dataset used is based in Bangalore, India. The dataset consists of 13000 records with 9 features. In the end, their findings within the restraints and techniques used, multivariable linear regression shows the best result in terms of accuracy and error.

Bansal et al. [2] study the comparison of two regression models in terms of performance. The two models chosen are multiple linear regression and linear regression with two specific scenarios in mind. First is to predict the salary of employees and second is to predict house prices. The author states that many factors are affecting the result. From the paper, all the features of a house are required to be included to yield an accurate representation. Their research uses a support vector machine (SVM) implemented with the aid of other Python libraries known as Pandas, Numpy, Matplotlib and Graphlab to achieve the result. They concluded that multiple linear regression gives a better performance than linear regression. The authors stated that it is better to use a bigger dataset size to produce a better understanding in the regression model used. The dataset used is based in Sydney and Melbourne, Australia and consists of 4600 records.

Mohd et al. [3] utilize machine learning techniques to predict house selling price using different regression models like random forest, decision tree, ridge, linear and lasso at Petaling Jaya town. Their finding indicates that usage of irrelevant features will reduce the accuracy of the prediction models which is why they proposed feature selection on the dataset to narrow down to important features. The initial dataset consists of 19 features. To reduce the number of features, the authors use the correlation matrix to drop the dependent features. In the end of the research, their finding concluded that random forest models yield the best result in terms of accuracy.

Mohd et al. [4] study the correlation between green building and house price. The authors use 5 different models which are linear, random forest, decision tree, lasso and ridge. The dataset used consist of 18 features. This paper stated that the usage of parameters tuning was applied to find the suitable parameters. Their finding suggested that random forest gives the best performance. Besides that, they also concluded that the

Table 1. Description of the features

Price	The price of the house
Location	The location of the house
Bedroom	The total number of bedrooms in the house
Bathroom	The total number of bathrooms in the house
Square feet size	The area size of the house in square feet
Price per room	The price per room

green building feature does not give much performance difference in any of the models implemented because of the weak correlation of green building to house price.

Kuvalekar et al. [5] study the prediction of market value of houses. The authors use the decision tree model to construct their prediction model. In their case, they use additional features like air quality and crime rate to predict house prices. The dataset collected is from the area of Mumbai. This paper stated that their decision tree model yields 89% accuracy in predicting the house prices.

Thamarai and Malarvizhi [6] managed to find the performance difference in two regression models used, which are multiple linear and the decision tree. Through their research, they suggested that to avoid unwanted rules in the decision tree model, the tree must be pruned. Their experiments are done using the Scikit-learn machine learning library. They concluded that multiple linear is better than the decision tree.

Manjula et al. [7] uses three regression models in order to predict housing prices. The three regression models are simple linear regression, multivariate regression and polynomial regression. The size of the dataset is 21,000 houses in the King County region in Seattle, United States of America. Simple linear regression has a very high error. Multivariate regression gives low error using together with only three features which are square feet size, number of bedrooms and number of bathrooms. Polynomial regression has a high error also because it tends to overfit. Overfit model can be reduced by using ridge regression.

3 Dataset

The dataset gathered from propertyguru.com, a popular property website available in Malaysia. The dataset gathered in the month of April 2022 is in Wilayah Persekutuan where it is a state in Malaysia. Collected raw dataset consists of 49,451 records.

The research uses a dataset that was captured from propertyguru.com. The records consist of 6 features that have the possibility of affecting the house prices. The features available on the dataset are described in the Table 1.

4 Research Methodology

The process that occur in this research are as follows:

- Data collection
- Data analysis
- Data preprocessing
- Training and testing models
- Performance metrics.

4.1 Data Collection

The dataset is critical to the project's starting. An appropriate dataset will produce a favorable result in the direction of the research goal.

4.2 Data Analysis

It is critical to analyze the data before applying any model techniques to it. As a result, it is necessary to examine the data and investigate the various aspects between them. Using data analysis, outliers in the data can be detected. Outliers usually occur when there are data collection errors and should be eliminated from the dataset to avoid noise.

4.3 Data Preprocessing

The practice of finding and correcting wrong data is known as data preprocessing. Two common data problems that need to be preprocessed are missing values and redundancies in the data. To achieve high prediction, it is best to rectify the data before deploying machine learning models.

The data modeling does not use the data that contains missing values. There is some data that has empty or missing values. For example, the column "Price", which indicates the current price of the house, has the missing price for the house. This information would impede the prediction result. Therefore, there is a need to mitigate any missing values in the selected features by eliminating them.

The nominal values of a feature that are too diverse will affect the results of the training and testing. A way to address this problem is by recategorizing to reduce the number of redundant nominal values. Recategorization of the nominal values will ensure that the populated values are in a condensed version. With recategorization, it will also make data visualization to be in a more manageable manner and the generated result has low prediction error.

4.4 Training and Testing Models

Machine learning techniques are applied here to train the model before testing. Data is separated into two parts which are training data and testing data. It is necessary to discover a good ratio between the two sets of these data. Separation into two data parts is also known as cross validation. Cross-validation is the process of running the model on the training data then using the generated model on the test data. The data is split into a 70:30 ratio. Training dataset is 70% while the test dataset is 30%.

The chosen machine learning techniques for training and testing are multiple linear regression, Bayesian ridge regression, decision tree and random forest regression.

4.4.1 Multiple Linear Regression

This multiple linear regression technique is more reliable. It works with many variables because it helps to estimate the value of an unknown behavior based on the known values of two or more properties.

4.4.2 Bayesian Ridge Regression

One of the most useful Bayesian regression techniques is Bayesian ridge regression which estimates a probabilistic model of the regression problem to predict continuous values. Bayesian ridge regression does not need to have any extra prior knowledge about the dataset.

4.4.3 Decision Tree Regression

This decision tree regression technique is deemed strong in handling outliers. Plus, the technique does not require normalization when handling features of data.

4.4.4 Random Forest Regression

Random forest is a powerful machine learning technique. It is very stable to a point where if given new data, it is not affected by it that much. Plus, it is not impacted by noises as compared to decision tree regression that uses a single tree.

4.5 Performance Metrics

To make a comparison among machine learning techniques, three performance metrics are used which are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2 score).

4.5.1 Mean Absolute Error (MAE)

The mean absolute error (MAE) is calculated by taking the averages of errors without the differentiations of plus or minus with the aid of absolute operations [8]. The average

over test data with the absolute differences between prediction and actual data. The MAE equation is stated below:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{1}$$

where n is the number of samples, y_i is the predicted value and x_i is the actual value.

4.5.2 Root Mean Squared Error (RMSE)

Root means squared error (RMSE) is a well-known equation for measuring the level of errors of a regression model [9]. Though, it is limited to only be used on models that produce errors in the same unit. The RMSE equation is stated below:

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

where n = number of samples in the data, x_i = actual value for the i-th sample and $\hat{x}_i$ = predicted value for the i-th sample. Generally the way it works is to get the predicted values to be deducted by the actual values, square them, sum up all of them, then divided by the number of samples and lastly squared root the entirety.

4.5.3 *Coefficient of Determination (R^2 Score)*

The coefficient of determination (R^2 score) is used to determine the relation between two input variables [10]. Another way of looking at R^2 would be the value 1 minus fraction differences between predicted values as numerator and mean values as denominator by the model. The R^2 score is from 0 to 1. The value 1 would indicate that the model is optimal while the value 0 would indicate that the model is not optimal at all. The coefficient of determination (R^2 score) equation is stated below:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y})^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2} \tag{3}$$

where n = number of samples, y_i = the actual value of the i-th sample, y_i = the predicted value of the i-th sample and $\bar{y}$ = mean of all samples.

The mean of all samples $\bar{y}$ equation is shown below:

$$\bar{y} = \frac{1}{n}\sum_{i=0}^{n-1} y_i \tag{4}$$

Table 2. Mean absolute error (MAE), root mean squared error (RMSE) and R^2 score results for four different machine learning techniques

Model	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	R^2 score
Bayesian ridge	158763.2906	299642.1369	0.8797
Multiple linear	158939.8088	299658.2138	0.8797
Decision tree	5719.5685	70260.2678	0.9933
Random forest	4439.3531	58758.2827	0.9953

5 Results

The Table 2 shows the mean absolute error (MAE), root mean squared error (RMSE) and R^2 score results for four different machine learning techniques applied in this research.

The research for this study is to find out performance scores from machine learning techniques in predicting house prices. The chosen machine learning techniques are Bayesian ridge regression, multiple linear regression, decision tree regression and random forest regression machine learning techniques.

With the aid of the machine learning performance results in the Table 2, the findings show that the random forest achieved the best R^2 score of 0.9953 followed by the decision tree which has the R^2 score of 0.9933. Both Bayesian ridge and multiple linear shared at the third place which have the same R^2 score of 0.8797. The high R^2 score from the random forest machine learning technique shows that it is able to predict house prices with a small margin of errors.

5.1 Data Visualizations of Results

The Fig. 1 shows the median price per square feet based on location in bar chart form. The bar chart depicted below shows that six expensive house prices are at Damansara, Desa Parkcity, Taman Tun Dr Ismail, Mont Kiara, KL city and lastly KL Sentral.

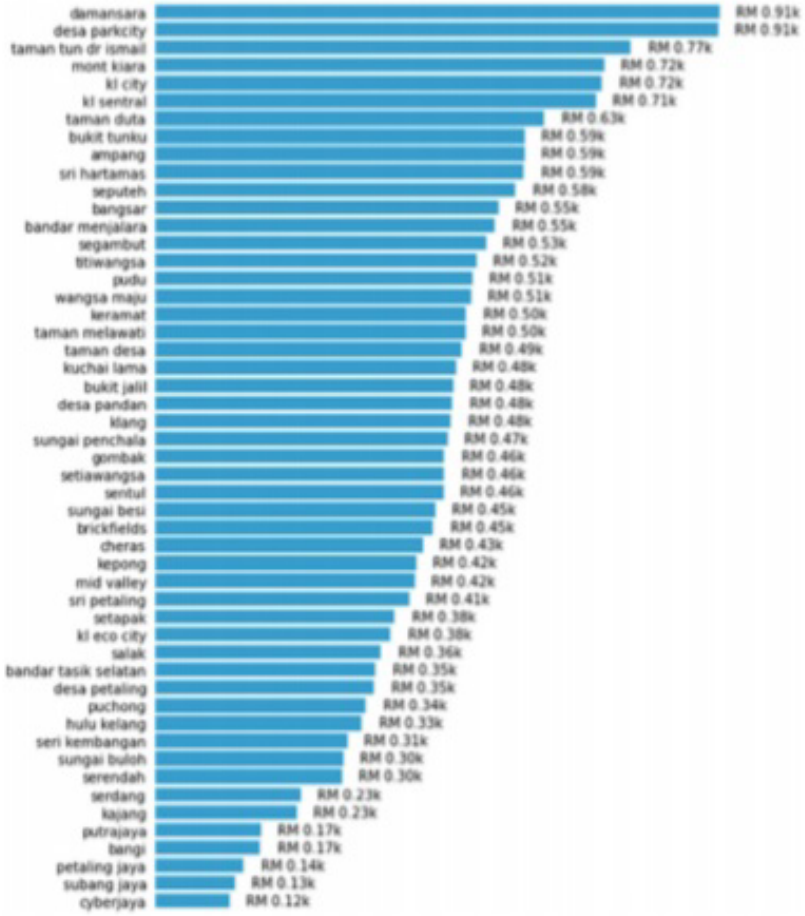

Fig. 1. Median price per square feet based on location.

The Fig. 2 shows the total number of available properties based on location in bar chart form. The bar chart depicted below shows that the biggest volume of houses is available in KL city.

The Fig. 3 shows the median property size based on location in bar chart form. The bar chart depicted below shows that the houses in Subang Jaya are having the biggest area size.

The Fig. 4 shows a snippet of first 10 records of actual value versus predicted value using the Bayesian ridge regression. Both predicted vertical bars and actual vertical bars having different heights indicates that the Bayesian ridge regression technique has a lower prediction score.

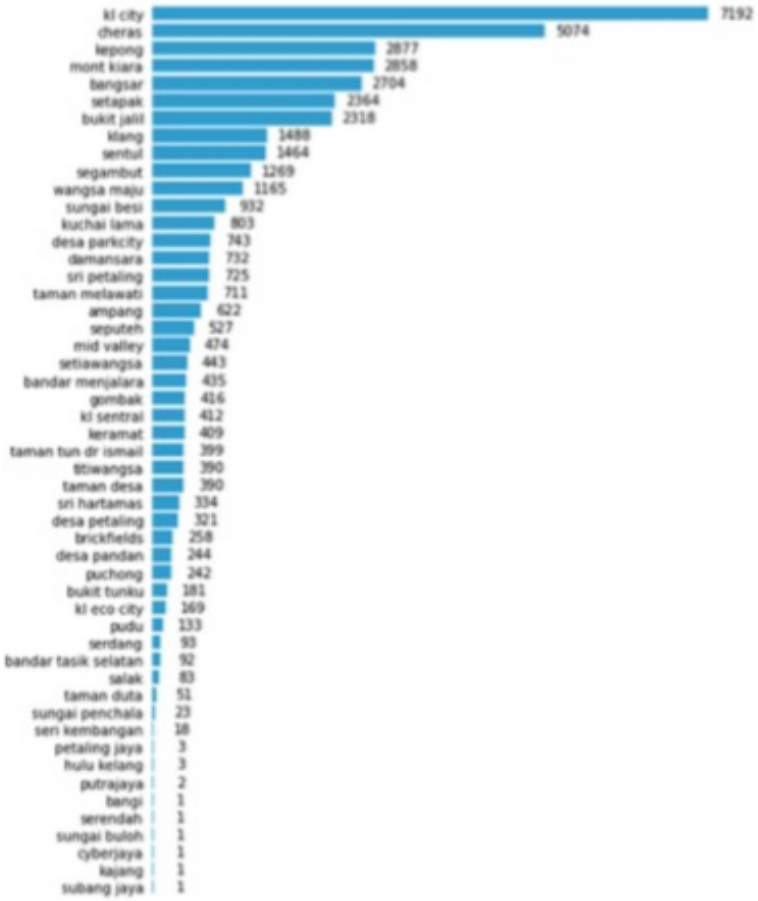

Fig. 2. Total number of available properties based on location.

The Fig. 5 shows a snippet of the first 10 records of actual value versus predicted value using the multiple linear regression. Both predicted vertical bars and actual vertical bars having the different heights indicates that the multiple linear technique has a lower prediction score.

The Fig. 6 shows a snippet of the first 10 records of actual value versus predicted value using the decision tree regression. Both predicted vertical bars and actual vertical bars having the same heights indicates that the decision tree learning technique has a high prediction score.

The Fig. 7 shows a snippet of the first 10 records of actual value versus predicted value using the random forest regression. Both predicted vertical bars and actual vertical bars having the same heights indicates that the random forest technique also has a high prediction score.

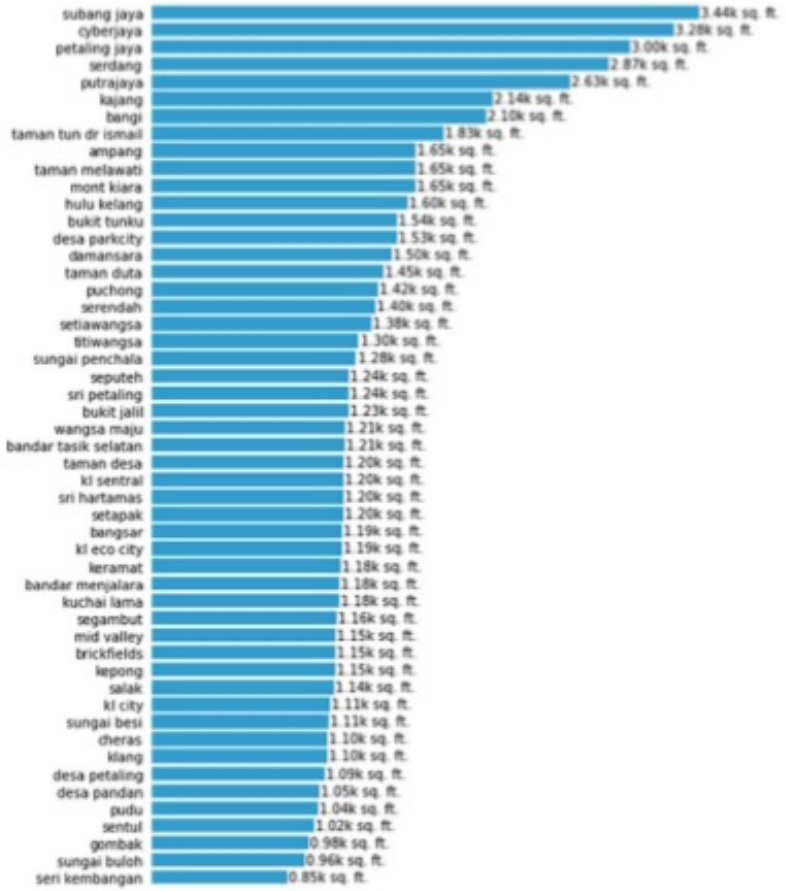

Fig. 3. Median property size based on location.

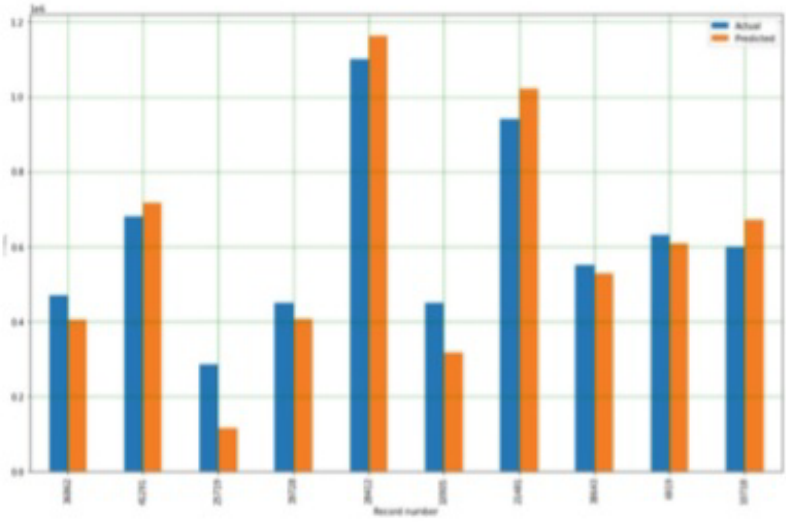

Fig. 4. Snippet of the first 10 records of actual value versus predicted value using the Bayesian ridge regression.

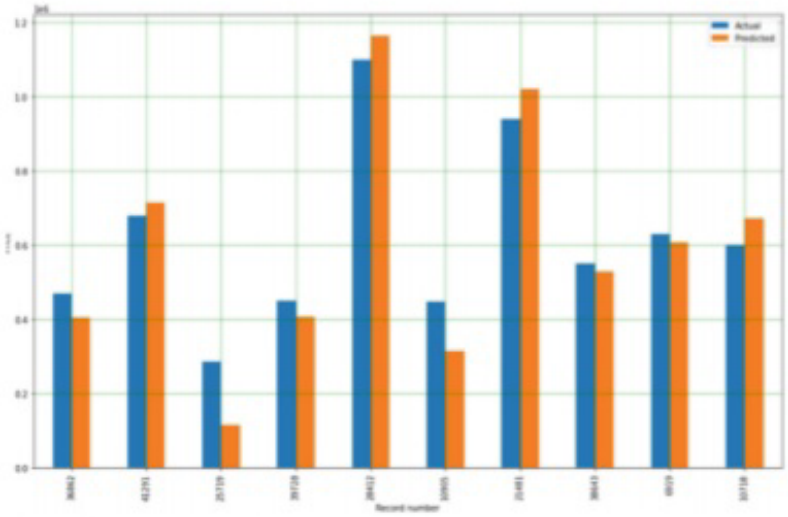

Fig. 5. Snippet of the first 10 records of actual value versus predicted value using the multiple linear regression.

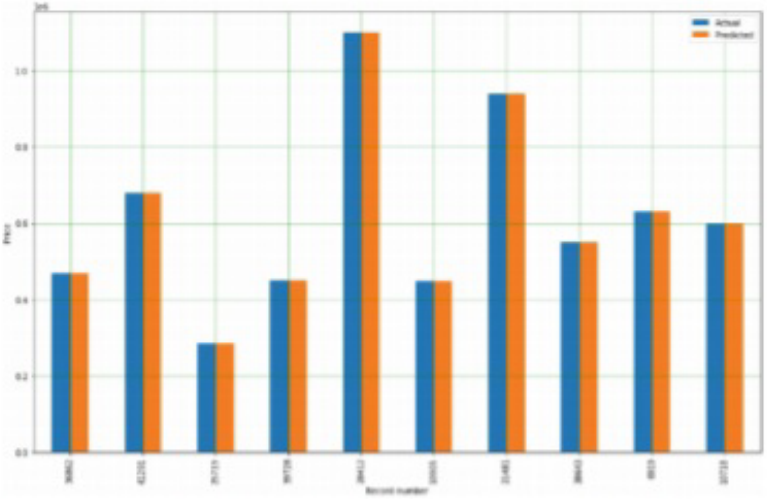

Fig. 6. Snippet of the first 10 records of actual value versus predicted value using the decision tree regression.

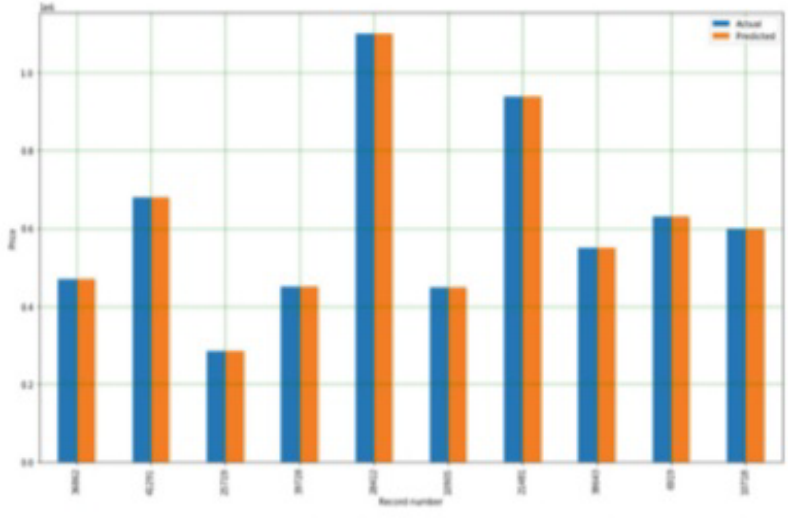

Fig. 7. Snippet of the first 10 records of actual value versus predicted value using the random forest regression.