ASSIGNMENT COVER SHEET

PROGRAMME	:	Master's in Business Analytics (MsBA)
SUBJECT CODE AND TITLE	:	BAA5073 – Data Mining
ASSIGNMENT TITLE	:	Data Mining Group Assignment – Predicting Housing Prices
LECTURER	:	Prof. Keshab Man Shrestha
ASSIGNMENT DUE DATE:		16/08/2024

STUDENT'S DECLARATION

1. I hereby declare that this assignment is based on my own work except where acknowledgement of sources is made.
2. I also declare that this work has not been previously submitted or concurrently submitted for any other courses in Sunway University/College or other institutions.
   [Submit "Turn-it-in" report (please tick √): Yes __√__ No _____]

NO.	NAME	STUDENT ID NO.	SIGNATURE	DATE
1.	Abdul Hakim Bin Kamalur Rahman	24015257	*Hakim*	09/08/2024
2.	Adrian Lim Phang Wei	16094088	*Adrian*	09/08/2024
3.	Harresh A/L Ragunathan	19076090	*Harresh*	09/08/2024
4.	Jaspreet Kaur A/P Surjit Singh	19122647	*Jaspreet*	09/08/2024

E-mail Address / Addresses (according to the order of names above):

1. 24015257@imail.sunway.edu.my	4. 19122647@imail.sunway.edu.my
2. 16094088@imail.sunway.edu.my	
3. 19076090@imail.sunway.edu.my	

APPROVAL FOR LATE SUBMISSION OF ASSIGNMENT (If applicable)

IF extension is granted, what is the revised due date? _____

Signature of Lecturer: _____ Date: _____

Marker's Comments:

Marks and / or Grade Awarded: _____ Date:_____

ADDENDUM

USE OF ARTIFICAL INTELLIGENCE (A.I.) DECLARATION

Students are allowed to use AI to support completion of assessments. However, students are reminded to do so ethically and transparently. This is so that (a) submissions can be fairly and accurately marked; and (b) feedback can be provided on the content that reflects student ability, in order to help with future submissions. Students are also reminded that in accordance with the University's Academic Malpractice Policy, Item 4.11.2, "*… the representation of work: written, visual, practical or otherwise, of any other person, including another student or* ***anonymous web-based material*** *[emphasis added], or any institution, as the candidate's own*" is considered malpractice.

Declaration

[√] I / We used the following A.I. tools to produce content in this submission:

Tool	**Purpose**	**Prompts**	**Sections where AI output was used / Outcome(s) in the submission**
e.g. ChatGPT	*e.g. Generating points for the essay* *Structuring the essay*	*e.g. "Give me 5 key talking points for an essay on…"* *"Show me a structure for an essay on…"*	*e.g. The main point for Section 1.2 and 1.3 were generated by AI, but the discussion was not.* *The organization / structure of the essay was suggested by AI*
e.g. Grammarly	*e.g. Correcting grammar and spelling, improving sentence structure*	*N/A*	*e.g. Grammarly suggestions were used for all sections of the essay*

Note: Add additional rows if necessary.

OR

[ ] I / We did not use any A.I. tools to produce any of the content in this submission.

NO.	NAME	STUDENT ID NO.	SIGNATURE	DATE
1.	Abdul Hakim Bin Kamalur Rahman	24015257	*Hakim*	09/08/2024
2.	Adrian Lim Phang Wei	16094088	*Adrian*	09/08/2024
3.	Harresh A/L Ragunathan	19076090	*Harresh*	09/08/2024
4.	Jaspreet Kaur A/P Surjit Singh	19122647	*Jaspreet*	09/08/2024

E-mail Address / Addresses (according to the order of names above):

1. 24015257@imail.sunway.edu.my	4. 19122647@imail.sunway.edu.my
2. 16094088@imail.sunway.edu.my	
3. 19076090@imail.sunway.edu.my	

Table of Contents

Abstract 4

1.0 Introduction 5

1.1 Research Problem 6

1.2 Research Objectives 6

1.3 Justification of Research 7

2.0 Literature Review 9

2.1 Machine Learning Comparative Review 9

2.2 Empirical Review 10

2.2.1 Housing Structure 10

2.2.2 Housing Conditions 11

2.2.3 Housing Environment 12

3.0 Data Preprocessing 13

4.0 Methodology 16

4.1 Model Testing 16

4.2 Evaluation Metrics 16

5.0 Results 19

6.0 Discussion 23

7.0 Limitations and Future Work 24

8.0 Conclusion 25

References 26

Appendix 30

Abstract

The growth of machine learning has become increasingly fast in this decade. Many applications and algorithms are evolving in machine learning day to day. One application that may be seen is house price prediction. Every year, house prices increase, which has necessitated the modelling of house price prediction. This study will investigate the issue of housing affordability in Malaysia, which has become a critical concern driven by rapid economic growth, high living standards, and fluctuating housing prices. As a developing nation, Malaysia's housing has evolved from necessities to key investment assets. As a result, many young and low-income households face hardships in achieving homeownership.

This study aims to improve the accuracy of housing price forecasts by using machine learning algorithms by focusing on the physical attributes, conditions, and environmental factors of houses. Using WEKA, it will delve into various types of machine learning algorithms, particularly regression-based models. At the end of this study, the top three models identified for predicting housing prices were SMO regression, Gaussian Processes, and Random Forest.

This study seeks to benefit a wide range of people including real estate investors, agents, construction companies, policymakers, and prospective home buyers. It also aims to help ease the process of decision-making and contribute to a more sustainable growth of Malaysia's housing industry.

1.0 Introduction

In recent years, Malaysia has concentrated its housing programs on eradicating poverty and integrating diverse ethnic communities. These programs seek to keep pace with the country's rapid economic growth and raise living standards. Housing in Malaysia stands as one of the most significant lifetime assets, transitioning from a fundamental necessity to a key investment as family income rises. However, housing affordability issues have been a subject of significant debate in Malaysia (Ling et al., n.d.). Such disruptions in the housing market are primarily due to the ongoing price fluctuations observed in Malaysian housing. Housing prices have increased due to various supply and demand factors in the housing industry, including land scarcity and labor shortages. As a result, many young and low-income households face significant challenges in obtaining homeownership (Beer et al., 2006; Berry and Dalton, 2004).

Globally, the economic burden of housing costs affects over 330 million households, with projections suggesting that by 2025, more than 1.6 billion people will reside in dangerous, low-quality, and financially constrained housing (Woetzel et al., 2014). With the average wage in Malaysia being approximately RM 6,338 per month, it is difficult for families to afford homes that cost more than RM 500,000 (Department of Statistics Malaysia, 2023). This disparity highlights the challenge for average households to keep up with rising housing prices, indicating a substantial gap between market forces and the housing demands of families and developers.

Due to the dynamic nature of Malaysia's housing markets, there is an increasing need for accurate housing price forecasting methods. The accuracy of these forecasts can be enhanced by using machine learning algorithms, which will benefit stakeholders including policymakers and homeowners. This study focuses on the determinants of housing prices in Malaysia that extend beyond the traditional economic variables, including specifically examining the impact of the physical attributes, conditions, and environmental factors of houses. Data visualization techniques and machine learning algorithms are employed to determine which algorithm exhibits the highest level of predictive accuracy.

1.1 Research Problem

This study intends to help the housing industry enhance the accuracy of determining housing prices based on the physical characteristics of the houses. Through the usage of data visualization techniques and machine learning algorithms, this study addresses several key questions:

1. What are the top three machine learning algorithms that are the most accurate in predicting housing prices among the analyzed algorithms?

2. How do certain housing structure features such as the overall area, number of bedrooms and bathrooms, number of stories, presence or absence of a basement, inclusion of a guest room, and the quantity of available parking space, affect housing prices?

3. How significantly do housing conditions, such as the availability of hot water heating and air conditioning systems, as well as the status of furnishings, affect housing prices?

4. To what extent does the housing environment, such as proximity to main roads and preference for particular areas, affect housing prices?

5. What are the factors that exert a significant effect on the housing price?

1.2 Research Objectives

This study aims to enhance understanding and contribute to the existing literature through the following approaches:

1. To identify the top three machine learning algorithms with the highest predictive accuracy by assessing various performance metrics.

2. To analyze the relationship between housing structure and prices, investigating how housing structure features such as the overall area, number of bedrooms and bathrooms, number of stories, presence or absence of a basement, inclusion of a guest room, and the quantity of available parking space, affect housing prices.

3. To assess the impact of housing conditions on prices, examining how factors like the availability of hot water heating and air conditioning systems, as well as the status of furnishings, affect housing prices.

4. To explore the relationship between housing environment and prices, investigating how proximity to main roads and preference for particular areas contribute to housing price variations.

5. To determine and evaluate the major factors influencing housing prices, taking into account a wide variety of variables and analyzing their individual and combined contributions to the pricing model.

1.3 Justification of Research

This study is vital as it is able to provide valuable insights for a wide array of stakeholders within the housing industry in Malaysia. This includes prospective home buyers, real estate investors, real estate agents, construction companies, as well as policymakers. This study enables prospective home buyers to better understand how housing structure impacts housing prices. Hence, this would allow them to make informed decisions that align with their financial and lifestyle goals.

This study provides real estate investors the ability to investigate the most effective machine learning algorithms for house price prediction. Before making an investment, they would be able to optimize their investing methods and make well-informed choices. In addition, Malaysian real estate agents can use these findings to improve the precision of their housing valuations, fostering transparency and trust. As highlighted by Rahadi et al. (2015), this study also focuses on the crucial yet often debated influence of housing structure on its market value. In order to develop accurate pricing models, key structural aspects are emphasized and thoroughly analyzed. This minimizes the possibility that houses would go unsold and guarantees that future housing developments will satisfy homeowners' long-term financial, and lifestyle needs.

Conversely, construction companies can use these insights obtained to plan and construct homes based on the needs and preferences of the current market. This would allow them to increase their

competitiveness within the housing industry. Moreover, policymakers can apply the knowledge gathered from this study to create measures that support the housing industry's sustainable growth while also attending to the needs of diverse communities. In summary, this study provides practical insights that can improve decision-making processes and allow the housing industry to become more efficient.

2.0 Literature Review

This section discusses and compares the machine learning algorithms utilized to identify the best algorithm that can produce highly accurate prediction outputs. An empirical review will also be conducted on housing variables such as the physical attributes, conditions, and environmental influences on housing prices. By examining these variables, a deeper understanding of their respective effects on the housing price dynamics can be achieved. These insights can aid the housing industry by strengthening the accuracy in predicting housing prices.

2.1 Machine Learning Comparative Review

In this study, seven machine learning algorithms were tested and compared using WEKA to identify the most optimal model for predicting housing prices. The algorithms tested include SMO regression, Random Forest, M5 Rules, Multilayer Perceptron, Linear Regression, M5P, and Gaussian Processes. Each model was evaluated based on its features and performance to determine the best fit for pricing houses.

SMO regression, a Support Vector Machine (SVM) algorithm, was utilized in this study. It is commonly used to handle tasks such as regression due to its ability to handle high-dimensional data and provide robust predictions (Schölkopf & Smola, 2020). Additionally, Random Forest was employed, which is commonly used for regression tasks. It is able to minimize overfitting tendencies and enhance prediction accuracy by aggregating multiple decision trees (Wang et al., 2021). Hence, Random Forest was also incorporated into our study.

Moving on, M5 Rules is a rule-based learning model that is ideal due to its interpretability and ability to apprehend non-linear relationships in data (Friedman et al., 2021). The strength of M5 Rules lies in its clear and understandable output, making data analysis straightforward. Additionally, Multilayer Perceptron (MLP), a type of artificial neural network, was utilized. This type of neural network is able to effectively handle complex models and patterns through non-linear functions and multiple layers (Goodfellow et al., 2020).

Linear regression, a common statistical model, was also utilized in this study due to its simplicity in capturing linear relationships within the dataset and its ease of interpretation (James et al., 2021). Since linear regression is very straightforward in nature, it is an ideal option for many real estate

applications. Gaussian Processes were also employed as they are useful for small datasets and non-linear regression. They are able to provide a probabilistic prospect to regression, offering uncertainty of measurement along with predictions (Williams & Rasmussen, 2020). Lastly, the M5P model used in this study is able to offer steady interpretability and prediction ability. combining decision trees and linear regression models (Quinlan, 2020).

In summary, SMO regression, Gaussian Processes, and Random Forest were selected as the top three models for housing price prediction in this study, due to their high correlation coefficients and lower prediction errors compared to the other models tested.

2.2 Empirical Review

2.2.1 Housing Structure

Research conducted in the past has intensively studied how housing structure characteristics can affect its value in the market. This study breaks down the concept of housing structure into seven tangible components which include the overall area, number of bedrooms and bathrooms, number of stories, presence or absence of a basement, inclusion of a guest room, and the quantity of available parking space (Wu & Sharma, 2018; Chiarazzo et al., 2014; McCord et al., 2012). These seven tangible components served as vital features of the houses. Hence housing structure, as suggested by Yang and Ghose (2019) has a significant impact on housing prices (H1).

Another vital factor is the overall housing area. As suggested by Su, Ceh, and Kim (2020), there is a positive relationship between the size of a house and its price. Buyers typically find larger homes more appealing since they usually offer greater living space and more features. Therefore, larger homes usually lead to higher prices in the market (H2). Additionally, houses with more bedrooms and bathrooms are commonly more appealing to buyers. Thus, adding extra bedrooms and bathrooms in a house will make it more convenient and comfortable for homeowners (Xu & Li, 2019). Hence, the number of bedrooms and bathrooms will impact housing prices (H3).

Furthermore, according to Zhang and Chen (2017), the number of stories in a home can affect housing prices. Multi-story houses commonly offer more living space and privacy, making them an optimal choice for various buyers. The value of multi-story houses increases with their utilization of land, particularly in urban areas where there is limited space. Besides, houses

featuring a guest room and basement will impact housing prices (H4). Having an additional basement and guest room not only serves as an enhancement to the appearance of a house but also serves as a functional space for some buyers in the market. Houses with features such as extra storage, living space, and guest rooms are desirable to buyers in the market (Han et al, 2021).

Consequently, another crucial factor impacting housing prices lies in the availability of parking spaces (H5). This is especially true in urban areas where parking spaces are limited. A house with ample parking space makes it appealing to buyers in the market who own several vehicles, which would ultimately influence housing prices (Smith & Miller, 2018). Generally, the presence of parking spaces improves the housing property's accessibility rather than merely making it more convenient.

H1: Housing structure has a significant influence on the rise in housing prices.

H2: Houses with a larger area will cause a rise in housing prices.

H3: Housing prices will rise when houses are equipped with more bedrooms and bathrooms.

H4: Houses furnished with a basement and guest room will cause a rise in housing prices.

H5: Larger parking spaces in houses will lead to an increase in housing prices.

2.2.2 Housing Conditions

A well-maintained property with the latest amenities is more attractive to buyers, as it requires minimal investment for further upgrades and repairs. Houses that are in good condition carry fewer risks and are more desirable, ultimately increasing their market value. As a result, (H6) housing conditions have a strong impact on housing prices (Bhardwaj & Shekhar, 2020).

A house that is fully equipped with air conditioning and water heater systems will cause a rise in housing prices (H7). Such houses that are equipped can be deemed more attractive and more valuable as they offer greater comfort and convenience to the buyers. Buyers are willing to pay more for houses that are furnished with the latest types of equipment and ready for move-in (Kim & Lee, 2020).

H6: There is a positive relationship between housing conditions and housing prices in the market.

H7: Houses that are equipped with water heaters and air-conditioning systems will cause a rise in housing prices.

2.2.3 Housing Environment

The environment of houses acts as another crucial factor in determining housing prices in the market. According to Jones and Brown (2023), the housing environment affects housing prices significantly (H8). This includes neighbourhood security and accessibility to essential amenities such as schools and workplaces, leading to a higher preference for that particular house among buyers. The added value to the living environment in such housing neighbourhoods will cause a rise in housing prices (Xie & Deng, 2021). A good housing environment enhances property values and the quality of life which makes them desirable for buyers to purchase.

H8: The housing environment affects housing prices.

3.0 Data Preprocessing

Table 1: Variables in the datasets

Variable Name	Description
area	The size of the house in square feet.
bedrooms	The number of bedrooms
bathrooms	The number of bathrooms
stories	The number of stories
main road	The location of the house, whether it is on the main road (yes, no)
guestroom	Whether there is a guest room (yes, no)
basement	Whether there is a basement (yes, no)
hot water heating	Whether there is hot water heating (yes, no)
air conditioning	Whether there is air conditioning (yes, no)
parking	The number of parking spaces
prefarea	Whether it is in a preferred area (yes, no)
furnishing status	The furnishing status (furnished, semi-furnished, unfurnished)
price	The price of the house

Data pre-processing is one of the important steps in data analysis and machine learning as it ensures the data used is cleaned from any unnecessary noise. Furthermore, it helps to make the data well-formatted, making it easier to use in model building.

In this assignment, the housing dataset is cleaned using the following steps:

1. Handling Missing Values

One of the factors that can impact the performance of machine learning models is missing value. Hence, this step is important as it reduces the bias in estimation and increases the efficiency of the machine learning model. In WEKA, the **'ReplaceMissingValues'** filter is used to handle missing data. By using this filter, it replaces the missing value in the dataset with the mean for the numeric

attributes and the most frequent value for nominal attributes. This filter will ensure that the dataset is complete with all the estimated values and ready for further analysis.

2. Removing Outliers and Extreme Value

The presence of outliers and extreme values will distort the results of data analysis which can lead to inaccurate models. In WEKA, the **'InterquartileRange'** filter identifies these values. The filter uses statistical techniques to calculate the outliers and extreme values. Once the values have been identified, it is then removed to improve the reliability and robustness of the dataset. This method also establishes a strong foundation for the dataset from the influence of anomalous data points.

3. Converting Nominal to Binary

Most machine learning algorithms often perform well when the data is numeric. Therefore, it is important to convert the data from nominal attributes to numeric or binary format. For this dataset, attributes such as 'mainroad', 'guestroom', 'basement', 'hotwaterheating', 'airconditioning', and 'prefarea' are categorical attributes (Yes, No). Hence, the **'NominalToBinary'** filter in WEKA is used to transform nominal attributes into binary attributes (0 = No and 1 = Yes).

4. Applying Dummy Variables

Some nominal attributes, like 'furnishingstatus' are converted to dummy variables. Dummy variables convert categorical data with binary values (0 and 1). The **'MakeIndicator'** filter is used to create dummy variables. This process creates additional columns for each category of the 'furnishingstatus' attribute (furnished, semi-furnished, and unfurnished) and classifies the feature as a binary value (0=No and 1=Yes). This conversion allows the dataset to be analyzed using models that require numeric input like regression while preserving the categorical information.

5. Normalizing The Data

Normalization converts the numeric attributes to a common range of [0 and 1]. This process helps to increase the accuracy by ensuring that no single attribute disproportionately influences the analysis due to its scale. In WEKA, the **'Normalize'** filter is used to obtain the common range. Before applying the filter, the class attributes must be removed first. Normalizing the data ensures that all the numeric attributes contribute equally to the analysis thus leading to more accurate and interpretable models.

6. Attribute Selection

In model building, it is important to choose the most relevant attributes as it can help reduce the dimensionality in the dataset, improve model performance, and reduce overfitting. The **'AttributeSelection'** in WEKA is used to perform this process. This step ensures that the attributes chosen will have the highest impact in predicting the target variable which is 'price'.

7. Setting 'Price' as the Class Attribute

In supervised learning, it is important to have a specific target variable which is known as the class attribute. In this housing dataset, 'price' is determined as the class attribute. By identifying the class attribute, it will help the machine learning algorithm to identify that 'price' is the variable that is used to predict based on the other attributes in the dataset.

4.0 Methodology

4.1 Model Testing

Once the data has gone through the preprocessing stage, the experimenting process may begin. During this process, WEKA Experimenter will be used to run the data through a handful of regression-based algorithms, in this case, “SMOreg”, “GaussianProcess”, “LinearRegression”, “MultilayerPerceptron”, “M5Rules”, “RandomForest”, and “M5P”. For the initial testing, these algorithms will be tested based on their default parameters in WEKA. These algorithms will then be evaluated using *k*-fold cross validation with 10 folds, as opposed to splitting the data into fixed training and testing sets (eg. 70-30 split). *K*-fold cross validation is a method of evaluating an algorithm’s accuracy by splitting the data randomly and equally into *k* subsets, called folds, and performing *k* iterations of testing (Kohavi, 1995). In this case, ten iterations will be performed where the data will be split randomly and evenly into ten folds. During each iteration, one fold will be used as the testing set, while the remaining nine will be used as the training set. By using this method, after ten iterations of *k*-fold cross validation, each fold will have been used at the testing set once. *K*-fold cross validation was chosen as the method of algorithm evaluation because, as opposed to the hold-out method, it allows all instances of the data to be used as the testing set. If the data is split into training and testing sets, the testing set will be fixed throughout the testing.

By using the Experimenter, the testing process for these algorithms is able to be automated, as WEKA will run 10 iterations of *K*-fold cross validation per algorithm chronologically, based on the instructed order. In this case, WEKA will run ten iterations of *K*-fold cross validation on the SMO regression algorithm. Once these ten iterations are complete, WEKA will then run ten iterations of *K*-fold cross validation on the Gaussian Processes algorithm. This testing process is continued until all seven algorithms have been evaluated. During this process, WEKA will output the results of each iteration of each algorithm into a “.arff” or “.csv” file. Once completed, the evaluation metrics of these algorithms are able to be compared simultaneously.

4.2 Evaluation Metrics

The metrics that will be analyzed include the Correlation Coefficient, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and the Root Relative

Squared Error (RRSE). A correlation coefficient is a measure that describes the strength of the relationships between attributes in a dataset (Shao, Zhuang, & Li, 2023). It is a measure that scales from -1 to 1; -1 indicates a perfect negative correlation between the variables, 1 indicates a perfect positive correlation between the variables, and 0 indicates no linear relationship between variables (Shao, Zhuang, & Li, 2023). MAE and RMSE are both metrics that evaluate the average model-performance error (Willmott & Matsuura, 2005). These two measures are similar; however, MAE is less sensitive to extreme values (Efron & Tibshirani, 1982).

Mean Absolute Error (MAE) is a measure that describes the magnitude of the difference between the predicted value of an observation and the actual value of that observation (Shao, Zhuang, & Li, 2023). It may be defined as:

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

Where,

- n = number of instances
- $\hat{y}_i$ = predicted value
- y_i = actual value

Root Mean Squared Error (RMSE) measures the distance between the data points and the regression line; measures how close the data is to the best fit line (Shao, Zhuang, & Li, 2023). It may be defined as:

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

Where,

- n = number of instances
- $\hat{y}_i$ = predicted value
- y_i = actual value

Relative Absolute Error (RAE) is a ratio of absolute errors produced by the predictive model to the errors produced by the actual model (Rao & Patel, 2021). It may be defined as:

$$RAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{\sum_{i=1}^{n}|y_i - \bar{y}|}$$

Where,

- $\hat{y}_i$ = predicted value
- y_i = actual value
- $\bar{y}$ = mean of actual values

Root Relative Squared Error (RRSE) is the square root of the relative squared error (RSE), which is a ratio that compares the error produced by the predictive model with the error produced by the actual model (Chu et al., 2021). It may be defined as:

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

Where,

- $\hat{y}_i$ = predicted value
- y_i = actual value
- $\bar{y}$ = mean of actual values

Generally speaking, a desirable model should have low values of MAE, RMSE, RAE, and RRSE (Kamani, Parmar, & Ghodasara, 2021).

5.0 Results

Table 2: Evaluation metrics of the tested models using K-fold cross validation (10 folds)

Model	**Correlation Coefficient**	**MAE**	**RMSE**	**RAE**	**RRSE**
SMO regression	0.8161	0.07361	0.10049	54.62%	58.26%
Gaussian Processes	0.8152	0.07448	0.10015	55.32%	58.15%
Linear Regression	0.8156	0.07490	0.10011	55.65%	58.16%
Multilayer Perceptron	0.7229	0.09910	0.14689	73.86%	85.83%
M5 Rules	0.8144	0.07497	0.10036	55.70%	58.30%
Random Forest	0.8144	0.07353	0.10128	54.51%	58.69%
M5P	0.8144	0.07497	0.10036	55.70%	58.30%

The evaluation metrics for each algorithm are retrieved and summarized into a table in order to make comparisons, as seen in Table 1. Based on the evaluation metrics in Table 1, it may be concluded that the top performing models are SMO regression, Random Forest, and Gaussian Processes. SMO regression emerges as the top performer among the other tested algorithms as it possesses the highest correlation coefficient of 0.8161. In terms of the metrics regarding errors, SMO regression produced one of the lowest MAEs of 0.07361, indicating it had a low average magnitude of errors. The produced RMSE, RAE, and RRSE using SMO regression was not the lowest among the tested models, nevertheless, these values are still close to the top performing models, allowing SMO regression to remain the best performing model. Following SMO regression, Random Forest is the next best performing model due to its high correlation coefficient of 0.8144. Additionally, among the tested models, Random Forest produced the lowest MAE of 0.07353, indicating it had the lowest average magnitude of errors. Similarly to SMO regression,

Random Forest's RMSE, RAE, and RRSE were not the lowest, however are low enough to compete with the best performing models. Lastly, following Random Forest, the next best performing model is Gaussian Processes due to its high correlation coefficient of 0.8152. However, its MAE, RMSE, RAE, and RRSE, although low enough to be competitive, are what holds this model back compared to the previous two models. Overall, these models are the best performing models due to their consistently high correlation coefficients and low error metrics.

Meanwhile, the remaining models did not perform to the standard of the best performing models due to their low correlation coefficients and high error metrics relative to the top three models. Namely, Multilayer Perceptron performed the worst, while Linear Regression, M5 Rules, and M5P were close contenders to the best performing models.

Table 3 – Evaluation metrics of the top 3 models after optimizations

Model	**Correlation Coefficient**	**MAE**	**RMSE**	**RAE**	**RRSE**
SMO regression	0.8164	0.07344	0.10070	54.48%	58.37%
Random Forest	0.8146	0.07270	0.10064	53.87%	58.29%
Gaussian Processes	0.8159	0.07494	0.10005	55.69%	58.12%

Although the top models perform well under the default settings and parameters, there is still room for improvements, which in turn, could provide a more accurate prediction model. In terms of the settings adjusted, SMO regression's 'c' parameter was adjusted from '1.0' to '0.1', Random Forest's 'maxDepth' was adjusted from '0' to '10', and Gaussian Processes' 'noise' was adjusted from '1.0' to '0.1'.

Overall, as seen in Table 1, Random Forest had improvements in all metrics, while SMO regression and Gaussian Processes only had improvements in three metrics; all three models had improvements in correlation coefficient. After optimizations, it is evident that SMO regression still produced the highest correlation coefficient, 0.8164, compared to the other models. However, it is noteworthy that the Gaussian Process' correlation coefficient had the biggest improvement, from

0.8152 to 0.8159. In terms of the metrics regarding error, Random Forest's MAE and RAE remain the lowest among the other two models, 0.07270 and 53.87% respectively. Meanwhile, Gaussian Process' RMSE and RRSE remained the lowest among the other two models, 0.10005 and 58.12% respectively.

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

price =

      0.2675 * area +
      0.2779 * bathrooms +
      0.137  * stories +
      0.0382 * mainroad +
      0.0304 * guestroom +
      0.0376 * basement +
      0.0833 * hotwaterheating +
      0.0828 * airconditioning +
      0.0811 * parking +
      0.0587 * prefarea +
     -0.038  * furnishingstatus=unfurnished +
      0.0404
```

Figure 1 – Attribute coefficients using Linear Regression

By using WEKA's Explorer feature, the coefficients of each variable may be analyzed from the test output of the Linear Regression model. Based on the output of this model, it is evident that the most significant factors that determine the price of housing are 'area', 'bathrooms', and 'stories'. This implies that if there are any increases in the size of the house, number of bathrooms, and/or number of stories, the price of the house would significantly increase. In terms of variables that have negative coefficients 'furnishingstatus=unfurnished' is the only variable that has a negative relationship with housing price, implying that an unfurnished house would likely lead to a decrease in price. Some notable variables regarding amenities that affect the price include 'hotwaterheating', 'parking', and 'airconditioning', indicating that houses that have water heating,

air conditioning, and/or have more parking spaces would likely lead to an increase in price. In terms of variables that slightly affect price, the variables include ‘mainroad’, ‘guestroom’, and ‘basement’, indicating that houses by a main road, that have a guestroom, and/or a basement would lead to an increase in price, but to a lesser degree than the previously mentioned amenities. Lastly, ‘prefarea’, houses in a preferred area, also increases the price, however, to a lesser degree than the most significant variables.

6.0 Discussion

With the results of the experiments, there are a handful of implications regarding the real estate industry that may be explored. Based on the results from the experiment, real estate companies or agents may want to consider implementing prediction models, specifically SMO regression, Random Forest, or Gaussian Processes, to predict housing prices in the future. These machine learning algorithms produced the best results among the seven tested algorithms, as demonstrated by their evaluation metrics. In short, these algorithms produced high correlation coefficients while maintaining overall minimal errors, justifying their significantly strong prediction performance. Real estate companies or parties who are interested in selling houses could use these prediction models to their advantage as it provides them an estimate of their property's worth based on features such as area, stories, and amenities. Having an on-demand tool that predicts housing prices could prove useful as it negates the need to physically scout prices of competitors' properties, which in turn may mitigate costs.

In terms of housing features that affect the price of housing, based on the experiment conducted, the three most significant features are the house's area, the number of stories, and the number of bathrooms. This provides the implication if the area of the house, the number of stories, and/or the number of bathrooms is higher, the price of the house would significantly increase, as these are all features regarding the structure of the house. Although these structural features significantly affect the price of housing, the house's amenities also affect the price, specifically houses with water heating, air conditioning, and parking spaces. Some other amenities that affect the price, however to a lesser degree, include houses with a guestroom and basement. In terms of the location of the house, a house in a preferred area would moderately increase the price, while a house by a main road would lead to an increase in price to a lesser degree. Lastly, the only housing feature that would lead to a decrease in price is if the house is unfurnished. Overall, structural features prove to be the most significant factor when it comes to increasing the price of housing. However, amenities such as water heating, air conditioning, and parking spaces should be taken into consideration as well.

7.0 Limitations and Future Work

To conduct the analysis, this study utilizes a dataset that thoroughly analyses a wide range of the physical characteristics of a house in order to accurately forecast housing prices. However, it is imperative to acknowledge that this study excludes important macroeconomic aspects that could have a substantial impact on the housing industry. These variables include gross domestic product (GDP) growth, employment levels, consumer confidence index, and inflation rates. These macroeconomic variables have the potential to impact consumers' ability to invest and make purchases which would impact housing prices and demand.

By excluding macroeconomic factors, this study aims to explore the direct relationship between housing prices and the physical attributes, conditions, and environmental factors of houses. This strategy is used to prevent any confounding effects that macroeconomic variables may cause. As a result, this makes it possible to analyze the inherent value of the non-economic housing characteristics more clearly. Despite this narrow focus, this study's scope is constrained because it cannot fully account for all the effects of the fluctuations in housing prices. In order to accurately predict housing prices, future research should take these macroeconomic aspects into account and incorporate them into a more thorough analysis.

To improve the reliability of the machine learning model, future research should expand beyond one geographic location. This can enhance the generalizability and robustness of the models. The dataset used in this study, retrieved from Kaggle, provides valuable insights specific to the area where the research was conducted. However, it has limitations that restrict the ability to make broad generalizations applicable to other regions.

In Malaysia, the factors that influence the price of the house may be different from Indonesia. Although they are neighbouring countries, factors such as economic conditions, demographic trends, and local policies will result in different prices even if the house is identical. Hence, having a multi-regional dataset will help to tackle this issue and improve the price prediction.

8.0 Conclusion

In conclusion, based on the findings from this study, the best algorithm to predict housing prices is SMO regression as it outputs the most desirable correlation coefficient along with producing low error metrics (MAE, RMSE, RAE, and RRSE) that are able to compete with the highest performing models. Furthermore, by using Linear Regression, real estate companies are able to analyze the significant factors that affect the price of houses. Based on the findings, the structural features, such as the area of the house, number of stories, and number of bathrooms affect the price the most. However, the house's amenities, such as water heating, air conditioning, and the amount of parking, also play a significant role in predicting the price. Lastly, an unfurnished house would lead to a decrease in price. Prediction models could prove to be useful to real estate companies as they will not only be able to predict the prices of houses based on their features, but also analyze how certain features would affect the price, providing them a general benchmark on how they should price their houses.

References

Beer, A., Kearins, B., & Pieters, H. (2006). Housing Affordability and Planning in Australia: The challenge of Policy under Neo-liberalism. *Housing Studies*, *22*(1), 11–24. https://doi.org/10.1080/02673030601024572

Berry, M., & Dalton, T. (2004). Housing prices and policy dilemmas: a peculiarly Australian problem? *Urban Policy and Research*, *22*(1), 69–91. https://doi.org/10.1080/0811114042000185509

Bhardwaj, N., & Shekhar, S. (2020). Impact of Housing Conditions on Property Values. *Real Estate Economics, 48*(2), 385-407.

Chen, Y., & Hao, Y. (2021). Comparative study of machine learning algorithms for housing price prediction. *Journal of Real Estate Research, 43*(3), 345-361.

Chiarazzo, V., Caggiani, L., Marinelli, M., & Ottomanelli, M. (2014). A Neural Network based model for real estate price estimation considering environmental quality of property location. *Transportation Research Procedia, 3*, 810-817.

Chu, D. T., Vu Ngoc, S. M., Vu Ngoc, H. T., Nguyen, T. N., Vu, T. M., Tran, B. H., Dinh, T. C., & Nga, V. B. (2021). An update on physical health and economic consequences of overweight and obesity. *Frontiers in Public Health, 9*, 729795. https://doi.org/10.3389/fpubh.2021.729795

Department of Statistics Malaysia. (2023, July 28). *Statistics on Household Income & Basic Amenities*. Retrieved July 23, 2024, from https://www.dosm.gov.my/portal-main/release-content/household-income-survey-report--malaysia--states

Efron, B., & Tibshirani, R. J. (1982). Statistical confidence of the empirical orthogonal functions. *Journal of Climate, 63*(12), 1309-1315. https://doi.org/10.1175/1520-0477(1982)063<1309

Friedman, J. H., Popescu, B. E., & Stein, C. (2021). Predictive learning via rule ensembles. *Journal of Machine Learning Research, 22*, 1-27.

Goodfellow, I., Bengio, Y., & Courville, A. (2020). *Deep learning*. MIT Press.

Han, J., Zhang, X., & Gong, X. (2021). The effect of basement and guestroom features on housing prices. *Housing Studies, 36*(4), 561-578.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer.

Jones, R., & Brown, J. (2020). The Influence of Neighborhood Environment on Housing Prices. *Urban Studies, 57*(5), 1023-1040.

Kamani, G. J., Parmar, R. S., & Ghodasara, Y. R. (2021). Machine learning models for productivity trend of rice crop. *Gujarat Journal of Extension Education, 32*(2), December 2021.

Kim, H., & Lee, S. (2020). The Value of Modern Amenities in Housing Markets. *Journal of Housing Economics, 47*, 101611.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)* (Vol. 2, pp. 1137-1143).

Ling, C. S., Almeida, S. J., & Wei, H. S. (n.d.). *Affordable Housing: Challenges and the Way Forward*. Bank Negara Malaysia. Retrieved July 23, 2024, from https://www.bnm.gov.my/documents/20124/770512/p3ba1.pdf

Ma, J., Deng, X., Chen, Y., & Li, Z. (2022). Real estate price prediction using machine learning algorithms. *Journal of Property Research, 39*(1), 1-20.

McCord, M., McCord, J., Davis, P. T., Haran, M., & McIlhatton, D. (2012). The impact of energy efficiency on housing prices in Northern Ireland. *International Journal of Housing Markets and Analysis, 5*(2), 170-190.

Quinlan, J. R. (2020). Learning with continuous classes. *Journal of Artificial Intelligence Research, 68*, 343-348.

Rahadi, R. A., Wiryono, S. K., Koesrindartoto, D. P., & Syamwil, I. B. (2015). Factors influencing the price of housing in Indonesia. *International Journal of Housing Markets and Analysis*, *8*(2), 169–188. https://doi.org/10.1108/ijhma-04-2014-0008

Rao, K. R., & Patel, H. P. (2021). Performance analysis of machine learning algorithms for prediction of chronic kidney disease. *International Journal of Computer Sciences and Engineering, 9*(9), 48-51. https://doi.org/10.26438/ijcse/v9i9.4851

Schölkopf, B., & Smola, A. J. (2020). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press.

Shao, J., Zhuang, S., & Li, X. (2023). Integrating data mining and machine learning for intelligent manufacturing systems. *Journal of Intelligent Manufacturing and Engineering Informatics, 100*, 100168. https://doi.org/10.1016/j.jjimei.2023.100168

Smith, T., & Miller, J. (2018). The role of parking facilities in urban property values. *Urban Planning and Development, 144*(1), 04017024.

Su, Y., Ceh, M., & Kim, S. (2020). House size and its impact on property prices: Evidence from metropolitan markets. *Regional Science and Urban Economics, 84*, 103599.

Wang, J., Liu, Y., & Zhang, S. (2021). Random forests: From fundamentals to state-of-the-art. *Artificial Intelligence Review, 54*(3), 2507-2531.

Williams, C. K. I., & Rasmussen, C. E. (2020). *Gaussian processes for machine learning*. MIT Press.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79-82.

Woetzel, J., Ram, S., Mischke, J., Garemo, N., & Sankhe, S. (2014, October). *A blueprint for addressing the global affordable housing challenge*. McKinsey Global Institute. Retrieved July 23, 2024, from https://www.mckinsey.com/~/media/mckinsey/featured%20insights/urbanization/tackling%20the%20worlds%20affordable%20housing%20challenge/mgi_affordable_housing_executive%20summary_october%202014.ashx

Wu, J., & Sharma, R. (2018). A critical review of hedonic price model applications in housing market research. *Journal of Housing and the Built Environment, 33*(2), 211-235.

Xie, J., & Deng, Y. (2021). Environmental quality and housing prices: An empirical analysis. *Journal of Real Estate Finance and Economics, 62*(1), 123-144.

Xu, L., & Li, Y. (2019). The impact of additional bedrooms and bathrooms on housing prices. *International Journal of Housing Markets and Analysis, 12*(3), 453-468.

Yang, Z., & Ghose, R. (2019). Structural determinants of housing prices: A comprehensive review. *Housing Policy Debate, 29*(4), 574-593.

Zhang, Y., & Chen, Z. (2017). The effect of building height on residential property prices. *Real Estate Economics, 45*(2), 319-350.

Appendix

```
Analysing:  Correlation_coefficient
Datasets:   1
Resultsets: 7
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       31/07/2024, 12:06 am


Dataset                   (1) functions | (2) functi (3) functi (4) functi (5) rules. (6) trees. (7) trees.
-------------------------------------------------------------------------------------------------------------
'Housing (cleaned 2)-weka(100)   0.8161 |   0.8152     0.8156     0.7229 *   0.8144     0.8114     0.8144
-------------------------------------------------------------------------------------------------------------
                                 (v/ /*) |    (0/1/0)    (0/1/0)    (0/0/1)    (0/1/0)    (0/1/0)    (0/1/0)


Key:
(1) functions.SMOreg '-C 1.0 -N 0 -I \"functions.supportVector.RegSMOImproved -T 0.001 -V -P 1.0E-12 -L 0.001
(2) functions.GaussianProcesses '-L 1.0 -N 0 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\" -S 1'
(3) functions.LinearRegression '-S 0 -R 1.0E-8 -num-decimal-places 4' -3364580862046573747
(4) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
(5) rules.M5Rules '-M 4.0 -num-decimal-places 4' -1746114858746563180
(6) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698
(7) trees.M5P '-M 4.0 -num-decimal-places 4' -6118439039768244417
```

Figure A1: Correlation coefficient

```
Analysing:  Mean_absolute_error
Datasets:   1
Resultsets: 7
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       31/07/2024, 12:08 am


Dataset                   (1) functions. | (2) functio (3) functio (4) functio (5) rules.M (6) trees.R (7) trees.M
-------------------------------------------------------------------------------------------------------------------
'Housing (cleaned 2)-weka(100)   0.07361 |   0.07448     0.07490     0.09910 v   0.07497     0.07353     0.07497
-------------------------------------------------------------------------------------------------------------------
                                 (v/ /*) |     (0/1/0)     (0/1/0)     (1/0/0)     (0/1/0)     (0/1/0)     (0/1/0)


Key:
(1) functions.SMOreg '-C 1.0 -N 0 -I \"functions.supportVector.RegSMOImproved -T 0.001 -V -P 1.0E-12 -L 0.001 -W 1\"
(2) functions.GaussianProcesses '-L 1.0 -N 0 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\" -S 1' -862006
(3) functions.LinearRegression '-S 0 -R 1.0E-8 -num-decimal-places 4' -3364580862046573747
(4) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
(5) rules.M5Rules '-M 4.0 -num-decimal-places 4' -1746114858746563180
(6) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698
(7) trees.M5P '-M 4.0 -num-decimal-places 4' -6118439039768244417
```

Figure A2: Mean absolute error

```
Analysing:  Root_mean_squared_error
Datasets:   1
Resultsets: 7
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       31/07/2024, 12:08 am


Dataset                  (1) functions. | (2) functio (3) functio (4) functio (5) rules.M (6) trees.R (7) trees.M
-----------------------------------------------------------------------------------------------------------------
'Housing (cleaned 2)-weka(100)   0.10049 |   0.10015     0.10010     0.14689 v   0.10036     0.10128     0.10036
-----------------------------------------------------------------------------------------------------------------
                               (v/ /*) |     (0/1/0)     (0/1/0)     (1/0/0)     (0/1/0)     (0/1/0)     (0/1/0)


Key:
(1) functions.SMOreg '-C 1.0 -N 0 -I \"functions.supportVector.RegSMOImproved -T 0.001 -V -P 1.0E-12 -L 0.001 -W 1\"
(2) functions.GaussianProcesses '-L 1.0 -N 0 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\" -S 1' -862006
(3) functions.LinearRegression '-S 0 -R 1.0E-8 -num-decimal-places 4' -3364580862046573747
(4) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
(5) rules.M5Rules '-M 4.0 -num-decimal-places 4' -1746114858746563180
(6) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698
(7) trees.M5P '-M 4.0 -num-decimal-places 4' -6118439039768244417
```

Figure A3: Root mean squared error

```
Analysing:  Relative_absolute_error
Datasets:   1
Resultsets: 7
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       31/07/2024, 12:09 am


Dataset                  (1) function | (2) funct (3) funct (4) funct (5) rules (6) trees (7) trees
-----------------------------------------------------------------------------------------------------
'Housing (cleaned 2)-weka(100)   54.62 |   55.32     55.65     73.86 v   55.70     54.51     55.70
-----------------------------------------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (1/0/0)   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) functions.SMOreg '-C 1.0 -N 0 -I \"functions.supportVector.RegSMOImproved -T 0.001 -V -P 1.0E-12 -I
(2) functions.GaussianProcesses '-L 1.0 -N 0 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\'
(3) functions.LinearRegression '-S 0 -R 1.0E-8 -num-decimal-places 4' -3364580862046573747
(4) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
(5) rules.M5Rules '-M 4.0 -num-decimal-places 4' -1746114858746563180
(6) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698
(7) trees.M5P '-M 4.0 -num-decimal-places 4' -6118439039768244417
```

Figure A4: Relative absolute error

```
Analysing:  Root_relative_squared_error
Datasets:   1
Resultsets: 7
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       31/07/2024, 12:09 am


Dataset                   (1) function |  (2) funct  (3) funct  (4) funct  (5) rules  (6) trees  (7) trees
-----------------------------------------------------------------------------------------------------------
'Housing (cleaned 2)-weka(100)   58.26 |    58.15     58.16     85.83 v    58.30      58.69      58.30
-----------------------------------------------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)    (0/1/0)    (1/0/0)    (0/1/0)    (0/1/0)    (0/1/0)


Key:
(1) functions.SMOreg '-C 1.0 -N 0 -I \"functions.supportVector.RegSMOImproved -T 0.001 -V -P 1.0E-12 -
(2) functions.GaussianProcesses '-L 1.0 -N 0 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\
(3) functions.LinearRegression '-S 0 -R 1.0E-8 -num-decimal-places 4' -3364580862046573747
(4) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
(5) rules.M5Rules '-M 4.0 -num-decimal-places 4' -1746114858746563180
(6) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698
(7) trees.M5P '-M 4.0 -num-decimal-places 4' -6118439039768244417
```

Figure A5: Root relative squared error

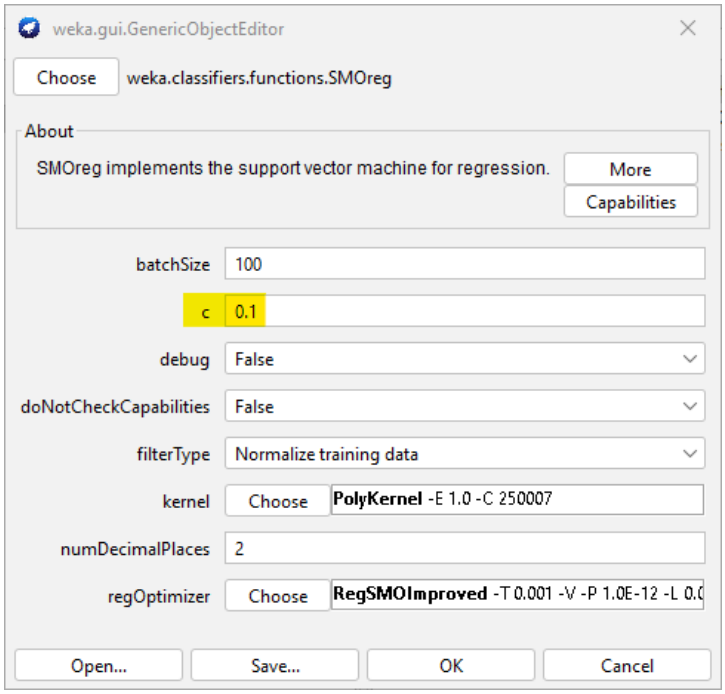

Figure A6: SMO regression optimized parameters

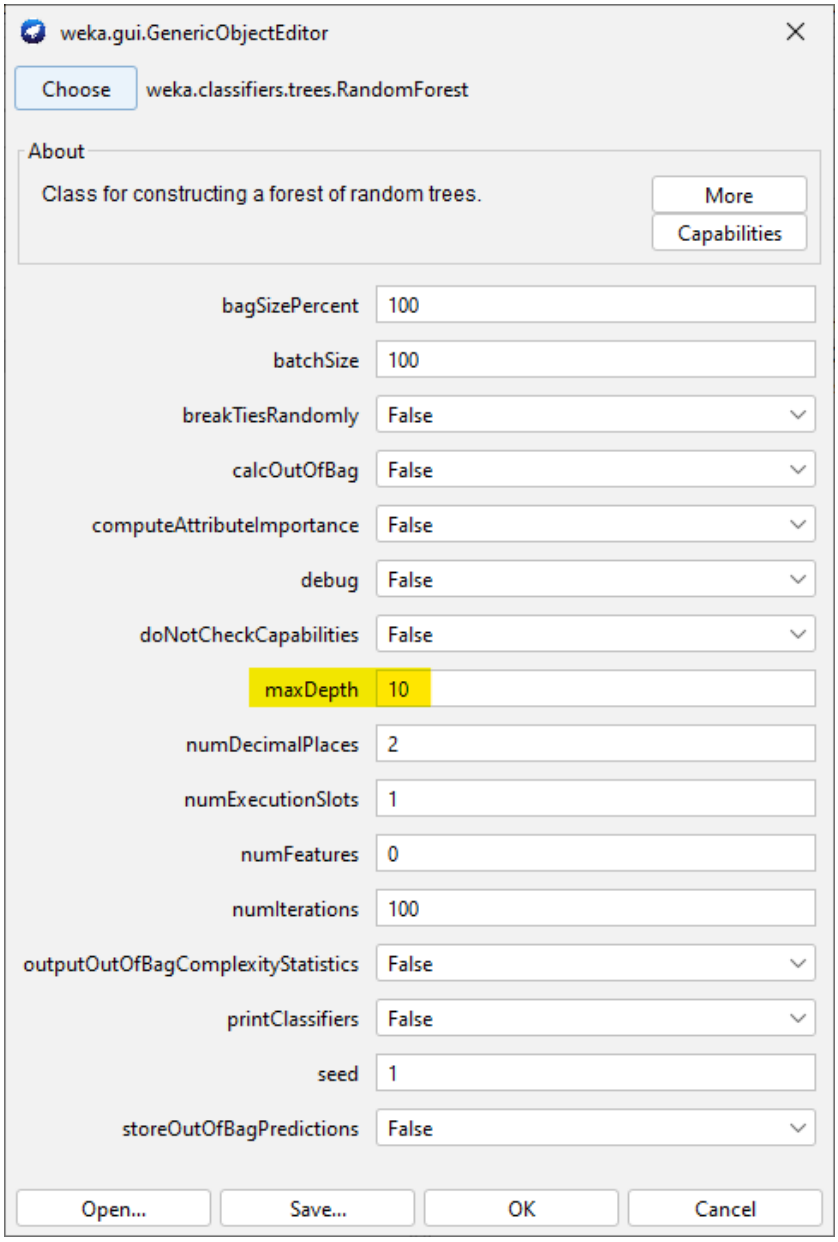

Figure A7: Random Forest optimized parameters

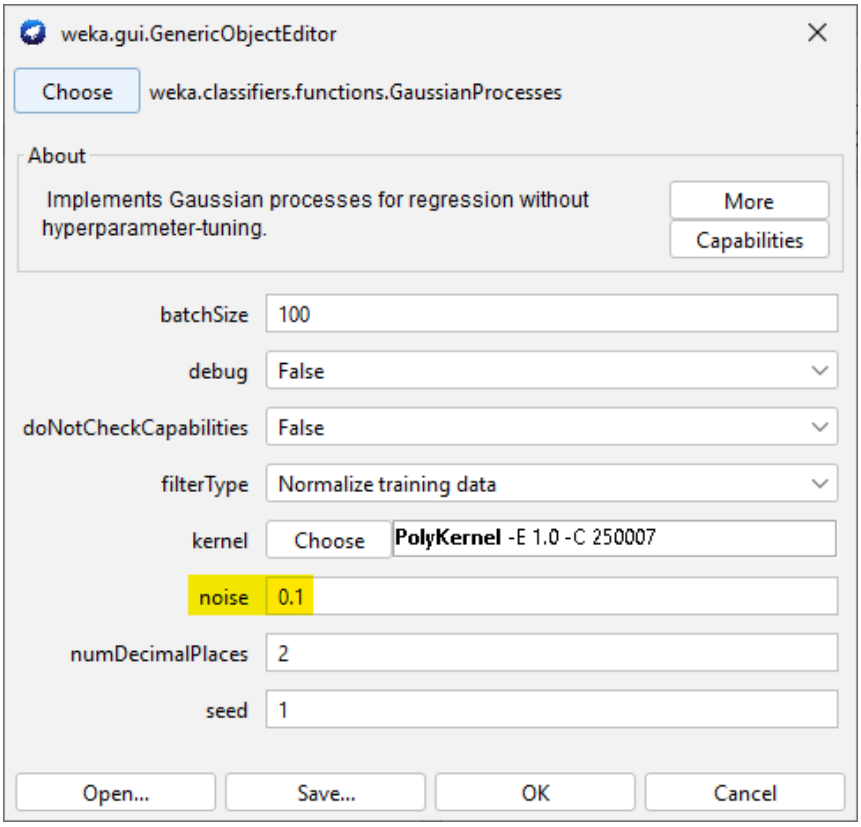

Figure A8: Gaussian Processes optimized parameters

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R
Analysing:  Correlation_coefficient
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       05/08/2024, 7:27 pm


Dataset                   (1) functions | (2) trees. (3) functi
----------------------------------------------------------------
'Housing (cleaned 2)-weka(100)   0.8164 |   0.8146     0.8159
----------------------------------------------------------------
                                (v/ /*) |    (0/1/0)    (0/1/0)


Key:
(1) functions.SMOreg '-C 0.1 -N 0 -I \"functions.supportVector.RegSM
(2) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.
(3) functions.GaussianProcesses '-L 0.1 -N 0 -K \"functions.supportV
```

Figure A9: Correlation coefficient (Optimized)

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -
Analysing:  Mean_absolute_error
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       05/08/2024, 7:28 pm


Dataset                   (1) functions. | (2) trees.R (3) functio
------------------------------------------------------------------
'Housing (cleaned 2)-weka(100)   0.07344 |   0.07270     0.07494
------------------------------------------------------------------
                                (v/ /*) |     (0/1/0)     (0/1/0)


Key:
(1) functions.SMOreg '-C 0.1 -N 0 -I \"functions.supportVector.RegSMOIm
(2) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001
(3) functions.GaussianProcesses '-L 0.1 -N 0 -K \"functions.supportVect
```

Figure A10: Mean absolute error (Optimized)

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S
Analysing:  Root_mean_squared_error
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       05/08/2024, 7:28 pm


Dataset                   (1) functions. | (2) trees.R (3) functio
------------------------------------------------------------------
'Housing (cleaned 2)-weka(100)   0.10070 |   0.10064     0.10005
------------------------------------------------------------------
                                (v/ /*) |     (0/1/0)     (0/1/0)


Key:
(1) functions.SMOreg '-C 0.1 -N 0 -I \"functions.supportVector.RegSMOImpr
(2) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -
(3) functions.GaussianProcesses '-L 0.1 -N 0 -K \"functions.supportVector
```

Figure A11: Root mean squared error (Optimized)

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1
Analysing:  Relative_absolute_error
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       05/08/2024, 7:29 pm


Dataset                   (1) function | (2) trees (3) funct
-----------------------------------------------------------
'Housing (cleaned 2)-weka(100)   54.48 |   53.87     55.69
-----------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)


Key:
(1) functions.SMOreg '-C 0.1 -N 0 -I \"functions.supportVector.R
(2) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -
(3) functions.GaussianProcesses '-L 0.1 -N 0 -K \"functions.supp
```

Figure A12: Relative absolute error (Optimized)

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -
Analysing:  Root_relative_squared_error
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       05/08/2024, 7:29 pm


Dataset                   (1) function | (2) trees (3) funct
-----------------------------------------------------------
'Housing (cleaned 2)-weka(100)   58.37 |   58.29     58.12
-----------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)


Key:
(1) functions.SMOreg '-C 0.1 -N 0 -I \"functions.supportVector.Reg
(2) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V
(3) functions.GaussianProcesses '-L 0.1 -N 0 -K \"functions.suppor
```

Figure A13: Root relative squared error (Optimized)