**Link to solution notebook:**

**Part:**

**2**

**Author**:

Enda Fan

25608579

**Date:**

2024.09.26

**Semester:**

1

**Year:**

2024

**University of Technology Sydney**

Table of Contents

[Table of Contents 2](#_Toc178260430)

[Summary or Executive Summary 3](#_Toc178260431)

[Introduction 4](#_Toc178260432)

[Report Body 5](#_Toc178260433)

[Methods/AI techniques used 5](#_Toc178260434)

[Implementation of AI techniques 6](#_Toc178260435)

[Comparisons of different AI techniques 7](#_Toc178260436)

[Conclusions and Recommendations 11](#_Toc178260437)

[List of References 13](#_Toc178260438)

[Individual reflection 14](#_Toc178260439)

## Summary or Executive Summary

Supervised learning has two problems, one is a classification problem, another is a regression problem. When the output is one of a ﬁnite set of values (such as sunny/cloudy/rainy or true/false), the learning problem is called classiﬁcation. When it is a number (such as tomorrow’s temperature, measured either as an integer or a real number), the learning problem has the (admittedly obscure 1) name regression (Russell, Stuart, & Peter Norvig, 2021).

The purpose of this project is to help potential homebuyers, developers, and investors make more informed decisions by applying multiple regression models to predict California home prices.

Home price e prediction is a complex task because home prices can be affected by a variety of factors, such as geographic location, housing characteristics, and economic conditions. It is the non-linear relationship between these factors that makes it challenging to accurately predict home prices. Characteristics in this dataset that may affect home prices include. MedInc: The median income of the area, which is usually positively correlated with home prices, with higher-income areas usually having higher home prices.2. HouseAge (House Age): The age of the home, which may affect home prices, with newer homes generally being priced higher than older homes, but in some cases, older homes may be more expensive due to location. 3. AveRooms (Average Rooms): The average number of rooms per home, usually reflecting the size of the home, more rooms mean higher home prices. 4. Latitude & Longitude: These are the coordinates of the geographic location, which directly affects the home price because location is a key factor in determining the price of a home. So, we can see that trying to predict house prices is very complicated and it is affected by many factors.

We used five regression models: support vector regression (SVR), linear regression, k-nearest neighbour (KNN), decision tree and random forest to evaluate their performance in house price prediction by R-squared、Mean Absolute Error (MAE)、Mean Absolute Percentage Error (MAPE)、Root Mean Squared Error (RMSE).

After experimenting with the data, we found that Random Forest and Decision Trees perform best in capturing complex non-linear relationships in the data. Whereas linear regression performs better when the data have strong linear relationships but performs poorly for complex data KNN is sensitive to local structure and is suitable for certain data with obvious local relationships. Finally, the comprehensive evaluation metrics show that Random Forest outperforms the other models’ overall performance.

For future research and practical applications, we should consider integrating multiple models and combining their respective advantages to improve prediction accuracy.

## Introduction

1. Problem and context:

This AT1 Example2 task was to apply an algorithm to solve a regression problem, which I chose the California Housing dataset from the website: <https://scikit-learn.org/stable/datasets/real_world.html#species-distribution-dataset>. Eight attributes were included: 1. MedInc: median income in a block group, 2. HouseAge: median house age in block group, 3. AveRooms: average number of rooms per household, 4. AveBedrms: average number of bedrooms per household, 5. Population: block group population, 6. AveOccup: average number of household members, 7. Latitude: block group latitude, 8. Longitude: block group longitude. I think people in our country are more concerned about such issues as house prices because, in our culture, a house and a car are both things that one must have to start a family, so I am also very interested in the issue of house price predictions. I hope that in the future if I have this data, I would also like to predict whether the house price will go up or down in my hometown in the coming period.

2. Why was it a problem:

House prices are affected by a variety of factors, including economic conditions, geographic location, and housing characteristics, and there is a complex non-linear relationship. Due to the diversity and uncertainty of these influencing factors, accurately forecasting house prices is a challenging task. Inaccurate house price forecasts could lead to home buyers paying excessive prices, investors taking increased risks, or banks taking unnecessary lending risks, and could even lead to market bubbles and financial crises.

3. Why was the project necessary:

Understanding and predicting changes in house prices is crucial for home buyers, sellers and property developers as it relates to financial decisions and investment strategies. Accurate house price forecasts not only help home buyers make more informed decisions in the market but also provide developers with pricing strategy references to avoid unnecessary risks in their investments. House prices are one of the most important indicators of the health of the economy. Accurate house price forecasts can help government departments and policymakers understand the development trend of the property market so that they can formulate more effective policies to regulate the property market and prevent the formation of price bubbles.

4. How was the problem solved:

This project predicted house prices by using a variety of regression algorithms such as Support Vector Regression, Linear Regression, K Nearest Neighbours, Decision Trees and Random Forests and then compared the performance of these models to find the optimal prediction method. Among the model performance measures include metrics such as R², MAE, MAPE and RMSE, which ensures the accuracy and reliability of the results.

## Report Body

### Methods/AI techniques used

1. Support Vector Regression, SVR: Think of SVR as wrapping a rubber band around a data point. It tries to find the smoothest line that encompasses most of the data points within the rubber band. This line represents the house price forecast, and SVR will try to keep the deviation of all data points from this line within a reasonable range. Of particular importance is the kernel function, which is a mathematical tool for mapping data in a low-dimensional space to a high-dimensional space, where it is easier to find a linear model that describes the relationships of the data. When predicting house prices, SVR first learns the relationship between features and house prices by using the features of the house and then finds an optimal line based on the features of the house to ensure that the predicted price is as close as possible to the actual price.

2. Linear Regression: Linear regression is a linear equation that represents the relationship between the characteristics and the target variable, like drawing a straight line with a straightedge that best fits the data, which is based on the location of the data points. The model adjusts the angle and position of the straight line so that it is as close as possible to all the data points. In house price forecasting, linear regression describes the relationship between features and price by using a straight line to describe, for example, the approximate increase in price for each square foot increase in the size of the house.

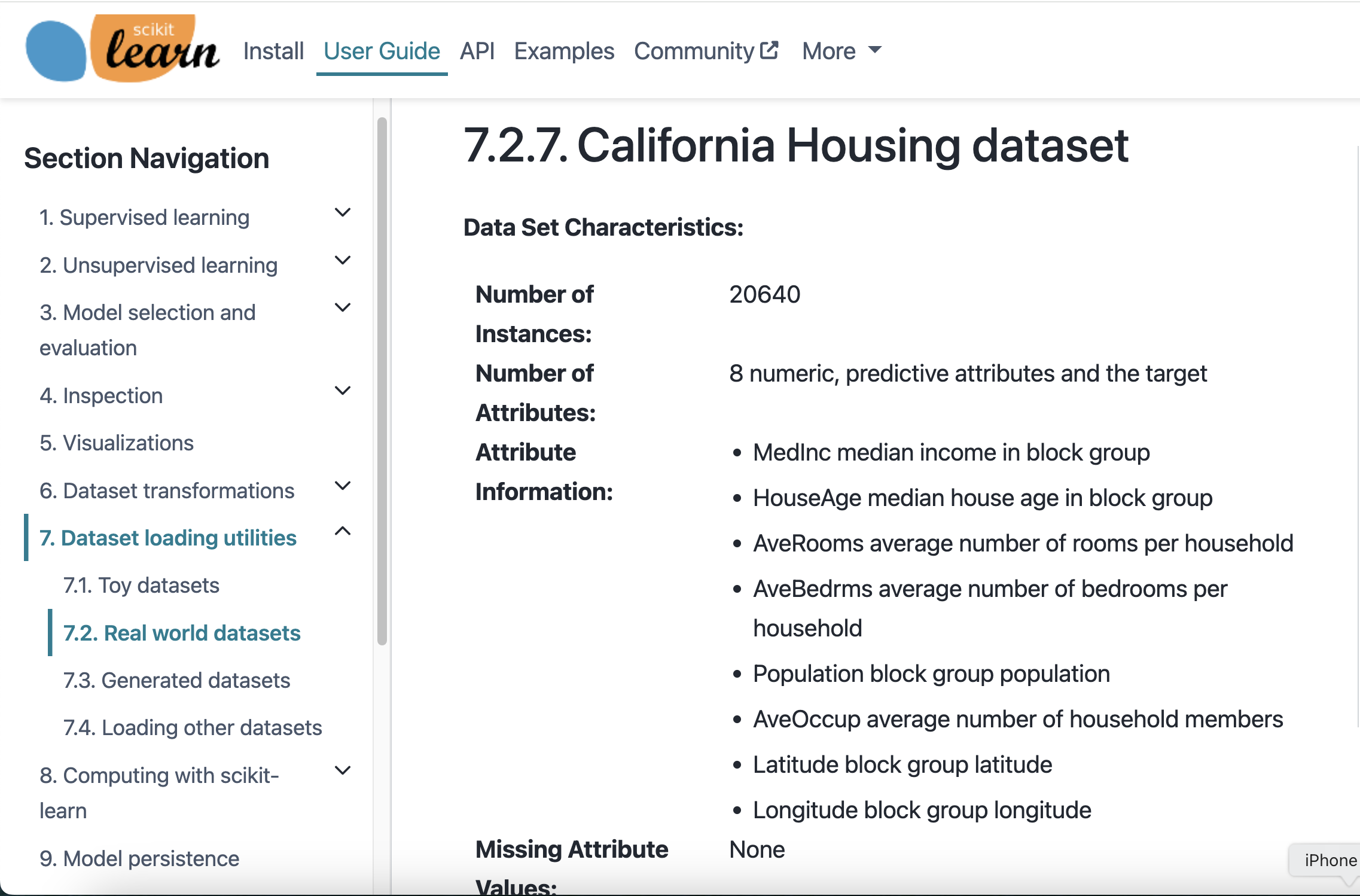
3. K-Nearest Neighbors Regression, KNN: For a sample to be predicted, the KNN looks for the K nearest neighbours that are most like it in the feature space, and then the average of the target values of these neighbours is used as the predicted value of the sample to be predicted. This process is like going to the neighbourhood. To predict the price of a house, the KNN looks for the houses most similar to it, sees how much they cost, and averages them. This way we know how much the house is worth. This method is like asking your neighbours about the price of a house, finding the house closest to you, and then making a prediction based on their price.

4. Decision Tree Regression: A decision tree forms a tree structure by recursively partitioning the data into different regions, like a binary tree in data structures, where each internal node represents a conditional judgement of a feature, and each leaf node represents a predicted value. The process is like a quiz game. It asks questions based on the characteristics of the house step by step, such as ‘Is the house larger than 100 square meters?’ ‘Does it have more than 3 rooms?’ The decision tree answers a series of questions, such as ‘Is the house larger than 100 square meters? A series of questions are answered to arrive at a predicted price.

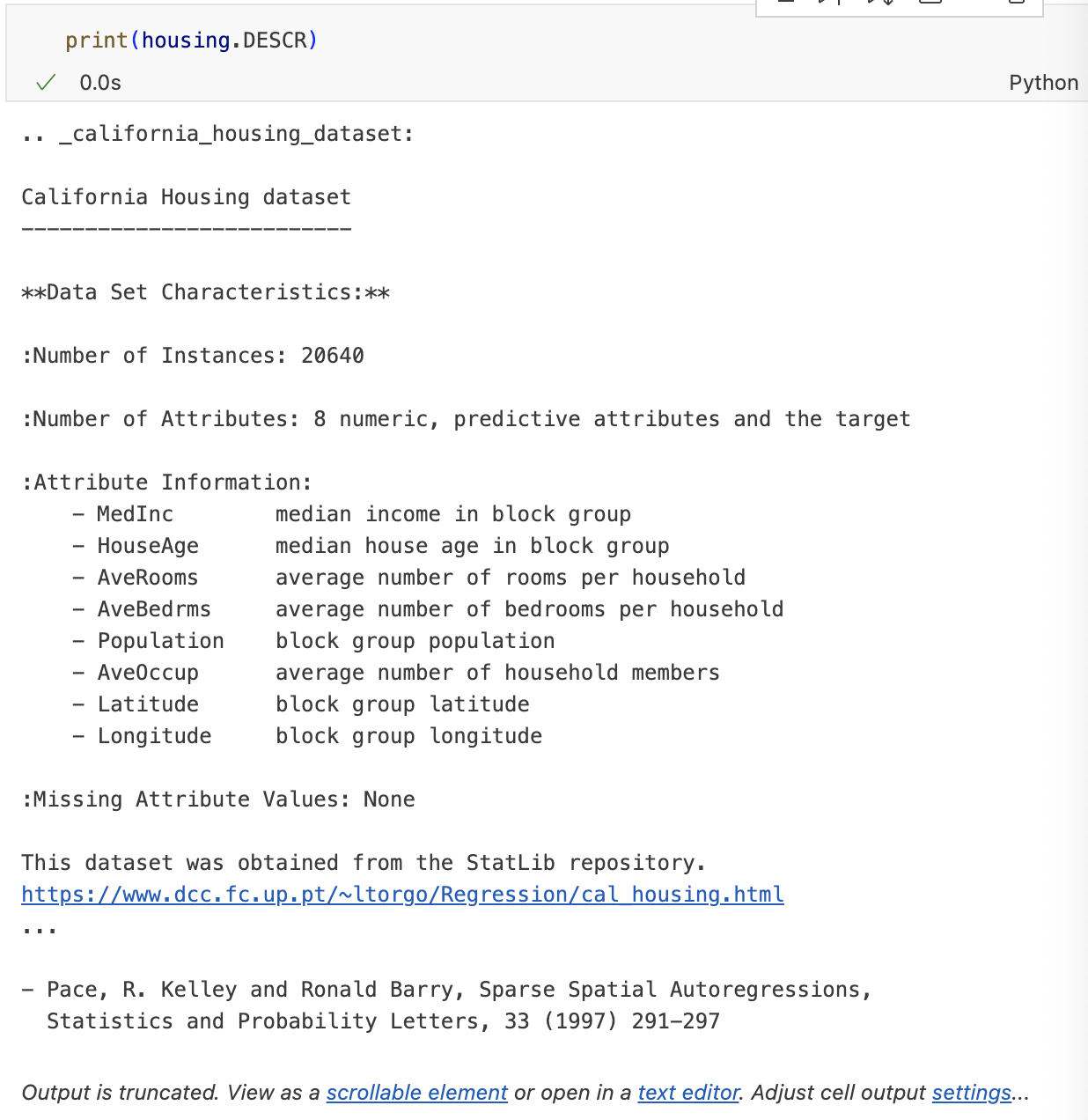
5. Random Forest Regression: Random Forest is an integrated learning method based on decision trees. It builds many different decision trees, each of which gives a price prediction, and then Random Forest averages the predictions from these trees as the final prediction. This improves the accuracy and stability of the prediction. In house price prediction, Random Forest trains multiple decision trees based on different combinations of features, each tree gives a price prediction and then votes to arrive at the final predicted price.

### Implementation of AI techniques

I used the California Housing dataset for home price forecasting.

*Figure 1 California Housing Dataset*

We imported the California Housing dataset from sklearn.datasets using the fetch\_california\_housing() function and stored it in the housing variable. The output from housing.data.shape (20640, 8) means that there are 20,640 samples in the dataset, each with eight features. housing.target.shape outputs (20640,) which mean that the target variable (house price) also contains 20,640 samples. This means that we have 20,640 different samples of housing data, each with eight input features, to predict the target price of the house. Again, with the output of housing.feature\_names, we can see what the properties of housing are.

*Figure 2 Details of the dataset*

‘DESCR’ is commonly used in Scikit-learn to provide descriptive information about a dataset, which contains a detailed textual description of the dataset. It allows the user to get the overall information about the dataset.

I also used three packages. The first one is the scikit-learn package, which is an open-source machine learning library based on Python. It contains the supervised learning algorithm of our learning, and the other two support packages are matplotlib and NumPy. Matplotlib is a Python library for data visualisation, mainly used for generating various types of charts, which can help us understand the relationships between data more intuitively. NumPy provides a rich set of mathematical functions. It is a fundamental tool for data science and machine learning. It helps me with data preprocessing, matrix calculations, and implementing various mathematical operations, such as calculating evaluation metrics such as RMSE, MAE, and MAPE.

 I used Matplotlib to plot the relationship between each attribute and house prices as a way of discovering the factors that influence house prices.

*Figure 3 Relationship between Median Income and house prices*

The image above shows the relationship between Median Income and home prices. The horizontal axis represents the median income for each neighbourhood and the vertical axis represents house prices. It can be seen that home prices trend upward as median income increases, indicating that the higher the income, the higher the home prices in the area. Therefore, it can be assumed that the amount of income affects house prices.

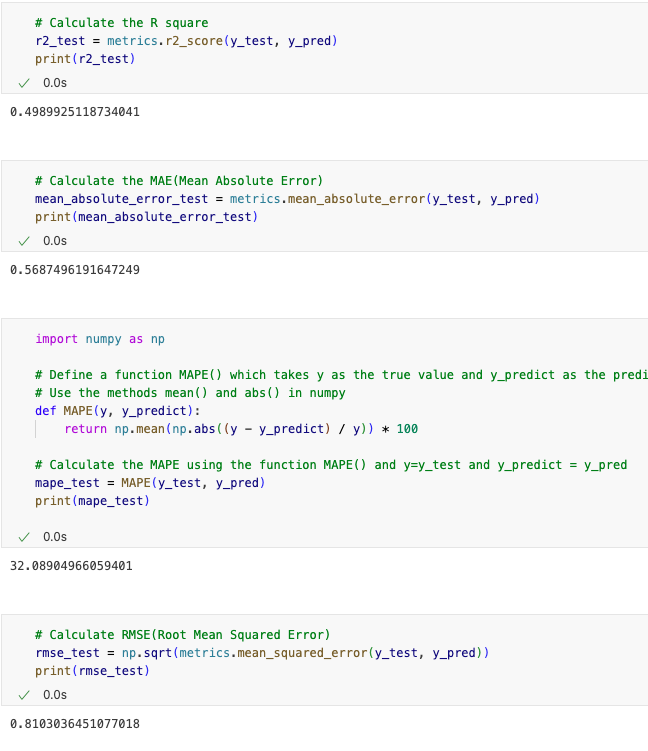
### Comparisons of different AI techniques

For these five algorithms, we use four metrics to evaluate their performance, which are R-square, MAE, MAPE, and RMSE.

1. R-squared (R²): It is also known as the coefficient of determination and is used to measure how well the model fits the data. R² indicates what proportion of the variation in the target variable can be explained by the model. It can range from 0 to 1, with the closer it is to 1, the better the model is at explaining the variation in the target variable. For example, R² = 0.8 means that the model explains 80 per cent of the variation in house prices.

2. Mean Absolute Error (MAE): It represents the average of the absolute value of the difference between the predicted and actual values and is used to measure the average level of model prediction error. MAE indicates how large the average error in model prediction is. For example, MAE = 3000 means that the average difference between the model's prediction and the actual house price is $3000.

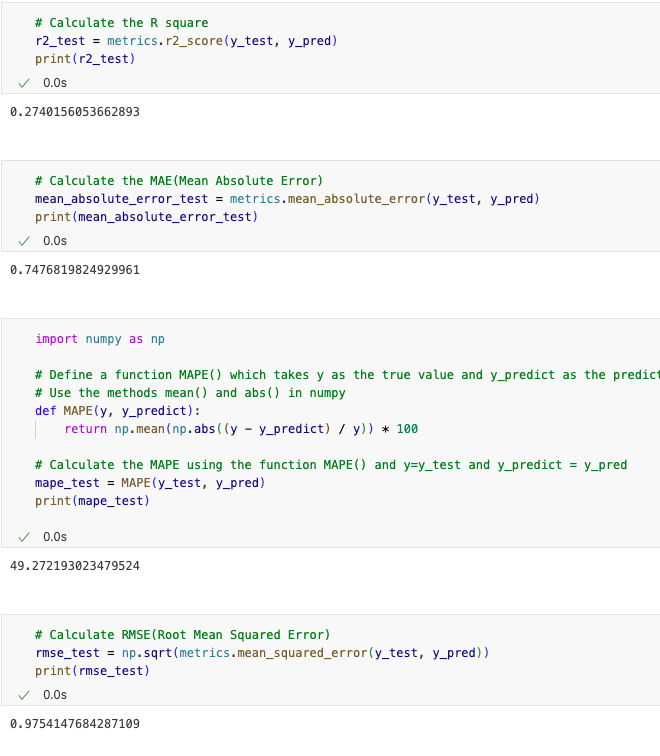
3. Mean Absolute Percentage Error (MAPE): It represents the average prediction error of the model as a percentage of the actual value. For example, MAPE = 10 per cent means that the model's predictions deviate on average by 10 per cent from the actual values.

4. Root Mean Squared Error (RMSE): It is the average of the squares of the difference between the predicted value and the actual value and then square it, which can be regarded as the standard deviation of the prediction error of the model, and the smaller the value, the more accurate the prediction of the model.

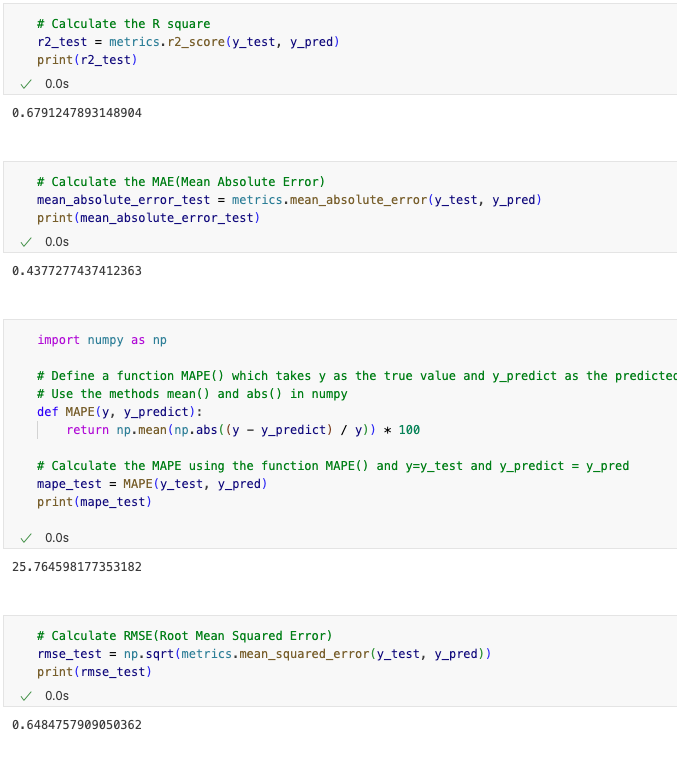
*Figure 4 The Result of SVR*

**

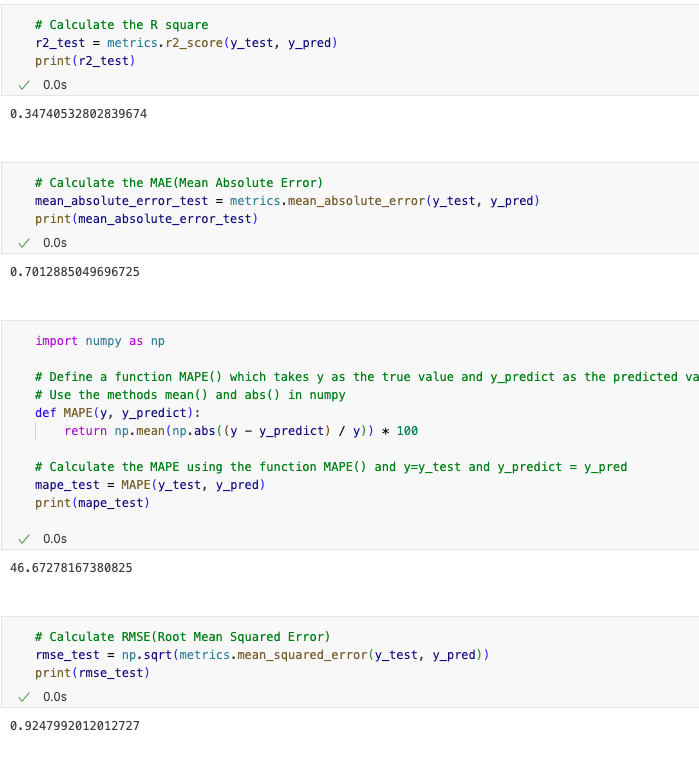
*Figure 5 The Result of Linear Regression*

**

*Figure 6 The Result of KNN*

**

*Figure 7 The Result of Decision Tree*

**

*Figure 8 The Result of Random Forest*

Based on these results, I want to make conclusions about them:

1. Random Forest has the highest R² with about 0.68, indicating that it is the best of the five models that can explain the data variation model. The lowest value is KNN with about 0.27, indicating that it does not capture the relationship between house prices and features well, and perhaps this dataset is not very suitable for using this model.

2. Random Forest has the smallest MAE, indicating that it has the smallest average error in predicting house prices, which means that the prediction results are very close to the actual values. And the value of KNN is the largest, which indicates that the error is also the largest, and there is a large deviation between the prediction and the actual house price.

3. Random Forest has the smallest MAPE of only 25.76%, indicating that it has the highest accuracy in predicting house prices. And the highest value remains KNN with 49.27%, indicating that it has the worst accuracy in predicting.

4. Random Forest has the smallest RMSE, with about 0.65, indicating that Random Forest performs the best among the five models and has the smallest prediction error. The KNN model has the largest prediction error, indicating that it is the least effective in predicting house prices.

## Conclusions and Recommendations

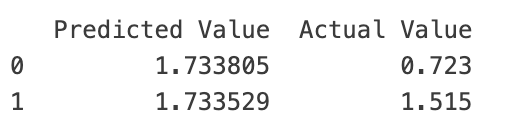
Support Vector Regression (SVR): Although the low score of R² indicates that it fails to explain the variations in the data well, its performance in MAE and RMSE is relatively stable, indicating that it is still relatively accurate in predicting some of the samples.

Linear regression: the R² score is slightly higher than SVR, indicating that it is slightly better at explaining the variation in the data. the MAE and RMSE also show that linear regression performs better on simple linear relationships but has limited performance on complex data.

K Nearest Neighbour (KNN): has the highest MAPE value, indicating that it is susceptible to extreme values and higher prediction errors when dealing with house price data. Although it has an advantage in capturing local relationships, its overall performance on the data is weak.

Decision Tree: The high R² score indicates that it is relatively good at interpreting the data, but the MAE and RMSE suggest that it may have produced large prediction errors for some samples, especially for those that did not appear in the training data.

Random Forest: performs best, suggesting that it is the most accurate predictor of house prices overall. This is attributed to its ability to combine the predictions of multiple trees, reducing the possibility of overfitting.



*Figure 9 Predicted Value of Random Forest*

Based on these experimental results, I suggest that in practical applications, the most suitable algorithm should be chosen according to the specific needs of the project and data characteristics. If the accuracy requirement is high and the computational cost is not considered, Random Forest must be a good choice; while in the case of limited computational resources, Linear Regression or Decision Tree may be a more cost-effective choice. Of course, we can also try to integrate several models by combining the advantages of multiple models to further improve the accuracy of prediction.

## List of References

Russell, Stuart, and Peter Norvig. (2021). Artificial Intelligence: a Modern Approach, Global Edition, Pearson Education Ltd.

## Individual reflection

I chose the five algorithms SVR, Linear Regression, KNN, and Decision Tree Random Forest for house price prediction, the reason I chose these five regression models is because I think these five algorithms are the most basic supervised learning algorithms, I should master the most basic ones first, and it is more beneficial for subsequent learning when I have my understanding. I chose to use the California Housing dataset and used specific evaluation metrics R², MAE, MAPE, and RMS as a way to measure the performance of the model.

Throughout the experiments I thought that tuning the parameters was a very complex problem: because I wanted them all to predict very accurate data, I was going to have to try to use different parameters, but because different models have different implementations, I

Through my experiments, I got a clearer understanding of the advantages and disadvantages of these five different algorithms. On the technical side, I learnt how to better select and adjust the parameters of the model and became more familiar with using Matplotlib for data visualization and analysis. On the personal side, I feel enhanced my patience and problem-solving skills.

If I had the chance to start again, I would try more algorithms, such as using RNN and LSTM, which should perform better. Or I would try a more systematic approach to tuning the parameters, which would make the model even better. In future projects, I would focus more on understanding the data, choosing more appropriate models, and paying more attention to the interpretation and application of evaluation metrics.