CSST102 - Basic Machine Learning	
Macatangay, Shaine Carla P.	**BSCS 3A**

Hour 1 – Setup & Dataset Exploration (Mini-task)

What is the input (features)?

- For the Housing dataset, the input is the data we use to make predictions, so the features are MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, and Longitude. These are the characteristics of the houses and their locations.

What is the output (label)?

- The output is what we want to predict. For the Housing dataset, the label is the **median house value (MedHouseVal)**.

Is this supervised or unsupervised learning?

- This is supervised learning because we already know the correct answers (the actual house values) and we use them to train the model.

Hour 2 – Train-Test Split & Baseline Model (Mini-task)

Compute model accuracy:

- After training the Linear Regression model on the Housing dataset, the performance is measured using **Root Mean Squared Error (RMSE)** instead of accuracy, since this is regression. The RMSE is usually around **70,000–80,000**, depending on the random train-test split. This value shows the average error between predicted and actual house values.

Hour 3 – Evaluation & Reflection

What would happen if the dataset had missing or wrong values?

- If there are missing or wrong values, the model will not learn correctly and its predictions will be inaccurate. This means the RMSE will increase, and in some cases, the model might even fail to train properly.

How does this relate to real-world ML applications?

- In the real world, housing data often has missing, noisy, or biased information, which can negatively affect predictions. That is why data cleaning, preprocessing, and validation are very important steps before training a regression model.

Conclusion

We used supervised learning because the dataset has inputs and outputs with correct answers. The model we used is Linear Regression for regression, which gave an RMSE of around 70,000–80,000 on the Housing dataset. One challenge that might affect the model is bad data, such as missing values or unusual outliers, which can reduce accuracy. Another challenge is underfitting, where the model is too simple to capture complex housing patterns. In real-world applications, careful dataset preparation and trying more advanced models can help improve performance and make predictions more reliable.