Joint Tech Internship Community Program

## Assignment 1: Machine Learning Terminologies

**Example Dataset:** **House Prices**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Size (sq ft) | Bedrooms | Age (years) | Location | Price ($) |
| 1 | 2000 | 3 | 10 | Urban | 500000 |
| 2 | 1500 | 2 | 5 | Suburban | 300000 |
| 3 | 2500 | 4 | 20 | Urban | 600000 |
| 4 | 800 | 1 | 15 | Rural | 150000 |
| 5 | 1200 | 2 | 5 | Suburban | 250000 |

**Explanation of Terminologies:**

**Feature**: Size (square ft), Bedrooms, Age (years), Location.

**Label**: Price ($).

**Prediction**: The estimated price generated by the model.

**Outlier**: If there was a house with a price of $1,500,000 in this dataset, it would be an outlier.

**Test Data**: A subset of this data not used for training, e.g., 1 or 2 records used to evaluate the model's performance.

**Training Data**: The subset of data used to train the model, e.g., 3 or 4 records.

**Model**: A regression model that predicts house prices based on features.

**Validation Data**: Another subset of data used to tune the hyperparameters, separate from training and test data.

**Hyperparameter**: Parameters like the learning rate, number of epochs, etc.

**Epoch**: One complete pass through all 3 or 4 training records.

**Loss Function**: Mean Squared Error (MSE) which measures the average squared difference between actual and predicted prices.

**Learning Rate**: A value that controls the step size during the optimization process.

**Overfitting**: If our model performs well on this dataset but poorly on new, unseen data.

**Underfitting**: If our model cannot capture the relationship between features and prices well.

**Regularization**: Techniques like L2 regularization to penalize large coefficients in the model.

**Cross-Validation**: Splitting the dataset into parts and training/testing on different partitions to ensure model reliability.

**Feature Engineering**: Creating new features from existing ones, e.g., combining Size and Bedrooms to create a new feature.

**Dimensionality Reduction**: Reducing the number of features, e.g., using only Size and Price.

**Bias**: Error from the model being too simplistic.

**Variance**: Error from the model being too complex and sensitive to training data.