**AI/ML Internship: Assignment 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **BHK** | **Rent** | **Area Locality** | **City** | **Floor** |
| **2** | **20000** | **Indiranagar** | **Bangalore** | **1** |
| **3** | **35000** | **Whitefield** | **Bangalore** | **3** |
| **1** | **15000** | **Kothrud** | **Pune** | **0** |
| **2** | **25000** | **Bandra** | **Mumbai** | **3** |
| **1** | **18000** | **Dwarka** | **Delhi** | **2** |

**1. Feature**

* A characteristic or property of the data that is used as input to a model.
* **Ex**: BHK, area, locality, city, floor.

**2. Label**

* The output or target variable that the model is trying to predict.
* **Ex**: Rent.

**3. Prediction**

* The output value predicted by the model.
* **Ex**: Predicted rent for a house.

**4. Outlier**

* A data point that is significantly different from other data points.
* **Example**: A house with a rent of 10,000 in a locality where the average rent is 1,000.

**5. Test Data**

* A subset of the data used to evaluate the performance of the model.
* **Ex**: A portion of the house rent dataset that the model has not seen during training.

**6. Training Data**

* The subset of data used to train the model.
* **Ex**: The portion of the house rent dataset used to build the model.

**7. Model**

* An algorithm that learns from data to make predictions or decisions.
* **Ex**: A linear regression model predicting house rent.

**8. Validation Data**

* A subset of data used to tune the model's hyperparameters.
* **Ex**: A separate portion of the house rent dataset used to adjust model settings.

**9. Hyperparameter**

* Parameters set before the learning process begins that control the learning process.
* **Ex**: Learning rate, number of epochs.

**10. Epoch**

* One complete pass through the entire training dataset.
* **Ex**: Training the model on the entire house rent dataset once.

**11. Loss Function**

* A function that measures how well the model's predictions match the true values.
* **Ex**: Mean Squared Error (MSE) for predicting house rent.

**12. Learning Rate**

* A hyperparameter that controls how much the model's weights are updated during training.
* **Ex**: A learning rate of 0.01.

**13. Overfitting**

* When a model learns the training data too well, including noise and outliers, and performs poorly on new data.
* **Ex**: A model that perfectly predicts rent for the training data but fails on test data.

**14. Underfitting**

* When a model is too simple to capture the underlying patterns in the data.
* **Ex**: A model that predicts the same rent for all houses regardless of their features.

**15. Regularization**

* Techniques used to prevent overfitting by adding a penalty to the loss function.
* **Ex**: L2 regularization (Ridge Regression).

**16. Cross-Validation**

* A technique to assess the model's performance by dividing the data into multiple folds and training/testing the model on these folds.
* **Ex**: 5-fold cross-validation on the house rent dataset.

**17. Feature Engineering**

* The process of creating new features or modifying existing ones to improve model performance.
* **Ex**: Creating a new feature "price per square foot" from rent and area.

**18. Dimensionality Reduction**

* Techniques to reduce the number of features while retaining important information.
* **Ex**: Principal Component Analysis (PCA) applied to the house rent dataset.

**19. Bias**

* Error introduced by approximating a real-world problem with a simplified model.
* **Ex**: Assuming rent only depends on BHK and ignoring other factors.

**20. Variance**

* Error introduced by the model's sensitivity to small fluctuations in the training data.
* **Ex**: A model predicting significantly different rents for similar houses due to noise in the training data.