Example:

Here is a sample test data of used house rent history.

Room ID	Building	Floor	Room Type	Square Feet	Occupied	Monthly Rent (USD)
101	Alpha	1	Studio	500	Yes	1500
202	Beta	2	One Bedroom	750	No	2000
303	Gamma	3	Two Bedroom	1000	Yes	2500
404	Delta	4	Studio	450	No	1400
505	Epsilon	5	One Bedroom	800	Yes	2200

Feature:

- **Definition**: An individual measurable property or characteristic of a phenomenon being observed.
- **Example**: In a dataset of houses, features might include the number of bedrooms, the size of the house, and the location.

Label:

- **Definition**: The output variable that the model is trying to predict.
- **Example**: In a dataset of houses, the label might be the price of the house.

Prediction:

- **Definition**: The output of a machine learning model after it has been trained on data.
- **Example**: A trained model predicts that a house with certain features will be priced at $300,000.

Outlier:

- **Definition**: A data point that differs significantly from other observations.
- **Example**: In a dataset of house prices, a house priced at $10,000,000 might be considered an outlier if most houses are priced between $100,000 and $500,000.

Test Data:

- **Definition**: A subset of the dataset used to evaluate the performance of the model.
- **Example**: After training a model on a dataset of house prices, you might use a separate 20% of the data as test data to see how well the model predicts house prices it has not seen before.

Training Data:

- **Definition**: The subset of the dataset used to train the model.
- **Example**: 80% of a dataset of house prices is used as training data to teach the model the relationship between features and prices.

Model:

- **Definition**: An algorithm or mathematical structure that makes predictions based on input data.
- **Example**: A linear regression model predicting house prices based on features such as size and location.

Validation Data:

- **Definition**: A subset of the dataset used to tune the model's hyperparameters and prevent overfitting.
- **Example**: 10% of the data is set aside to validate the model's performance during training and adjust parameters like learning rate.

Hyperparameter:

- **Definition**: Parameters that are set before the learning process begins and control the training process.
- **Example**: The learning rate in a neural network.

Epoch:

- **Definition**: One complete pass through the entire training dataset.

- **Example**: In training a neural network on house prices, one epoch means the model has seen all houses in the training data once.

Loss Function:

- **Definition**: A function that measures how well the model's predictions match the actual data.
- **Example**: Mean Squared Error (MSE) is a common loss function used for regression tasks.

Learning Rate:

- **Definition**: A hyperparameter that controls how much the model's weights are adjusted during training.
- **Example**: A learning rate of 0.01 might be used to control the step size in gradient descent.

Overfitting:

- **Definition**: When a model learns the training data too well, including the noise, and performs poorly on new data.
- **Example**: A model that predicts training house prices perfectly but fails to generalize to unseen houses.

Underfitting:

- **Definition**: When a model is too simple to capture the underlying pattern of the data.
- **Example**: A linear model predicting house prices might underfit if the relationship between features and prices is nonlinear.

Regularization:

- **Definition**: Techniques used to prevent overfitting by adding a penalty for larger model weights.
- **Example**: L2 regularization (Ridge Regression) adds a penalty equal to the sum of the squared coefficients.

Cross-Validation:

- **Definition**: A technique for assessing how the model will generalize to an independent dataset.
- **Example**: K-Fold Cross-Validation involves splitting the data into K parts and training/testing K times, each time using a different part as the test set.

Feature Engineering:

- **Definition**: The process of using domain knowledge to create new features that make machine learning algorithms work better.
- **Example**: Creating a new feature "price per square foot" from the existing features "price" and "size" of a house.

Dimensionality Reduction:

- **Definition**: Techniques to reduce the number of features in a dataset while retaining important information.
- **Example**: Principal Component Analysis (PCA) reduces the dimensionality of data by transforming it to a new set of variables.

Bias:

- **Definition**: The error introduced by approximating a real-world problem, which may be complex, by a simplified model.
- **Example**: A high bias model might assume that house prices are solely determined by house size, ignoring other factors.

Variance:

- **Definition**: The model's sensitivity to the specific training data it was trained on.
- **Example**: A high variance model may accurately predict house prices for the training data but performs poorly on new data because it has learned noise in the training data.