Task 8: – Machine Learning

1. What are Overfitting and Underfitting?

Overfitting

Overfitting occurs when a machine learning model learns the training data too deeply, including noise, outliers, and irrelevant patterns. As a result, the model performs extremely well on the training dataset but poorly on new or unseen data. This usually happens when the model is too complex, such as having too many parameters or being trained for too long. Overfitting reduces a model's ability to generalize and can be controlled using techniques like regularization, cross-validation, and reducing model complexity.

Underfitting

Underfitting happens when a model is too simple to capture the underlying structure or patterns in the data. In this case, the model performs poorly on both the training and testing datasets because it fails to learn meaningful relationships. Underfitting is often caused by using overly simple models, insufficient training time, or lack of relevant features. Improving model complexity, adding more features, or training longer can help reduce underfitting.

2. What is the difference between Supervised and Unsupervised Learning?

- **Supervised Learning** uses labeled data, meaning the input data is paired with correct output values. The model learns by mapping inputs to known outputs and improves its accuracy by minimizing errors during training. It is commonly used for tasks such as classification and regression, including spam detection, disease prediction, and house price estimation.

- **Unsupervised Learning** works with unlabeled data and focuses on identifying hidden patterns or structures within the data. Instead of predicting a specific outcome, the model groups similar data points or reduces complexity. Typical applications include customer segmentation, clustering, market basket analysis, and dimensionality reduction.

3. What is a training dataset and a testing dataset? Why is data splitting important?

Training Dataset

A training dataset is the portion of the data used to teach a machine learning model how to recognize patterns and relationships between input features and the target variable. During training, the model adjusts its parameters to minimize prediction errors. The quality and size of the training dataset play a critical role in determining how well the model learns.

Testing Dataset

A testing dataset is a separate portion of the data used to evaluate the performance of a trained model on unseen data. It helps measure how well the model generalizes beyond the training data. Data splitting is important because it prevents the model from simply memorizing the data and provides a realistic estimate of how the model will perform in real-world scenarios.

4 .What is feature scaling and why is it needed in some algorithms?

Feature scaling is the process of transforming input features so that they lie within a similar range or scale. This is especially important for machine learning algorithms that rely on distance calculations or gradient optimization. Without feature scaling, features with larger numerical values can dominate the learning process and negatively affect model performance. Common feature scaling techniques include normalization and standardization.

5. How does a Linear Regression model work?

Linear Regression works by modeling the relationship between input features and a continuous target variable using a straight line. It finds the best-fit line by minimizing the error between predicted and actual values, usually using a method like least squares. The model predicts output by multiplying input features with learned coefficients and adding a bias term.