MACHINE LEARNING

ASSIGNMENT 1

Sample data for crop yield:

YEAR	CROP	AREA (hectares)	FERTILIZERS USED (kg/hectare)	YIELD (tons/hectare)
2020	Wheat	100	50	1
2020	Corn	150	55	2.5
2021	Rice	45	20	1
2021	Rice	70	50	1.5
2022	Wheat	23	10	0.5

The Terminologies used:

1.Feature

Feature of this datasets are year, crop, area, fertilizers.

2. Label

The yield (tons/hectare) of the crop.

3.Prediction

Predicting the yield per hectare based on the input such as crop, area, fertilizers used.

4.Outlier

The yield of the crop is compared with the other crops yield.

5.Test data

The set of crop yield data used for testing the model.

6.Training model

The yield of the particular crop is collected for every year to train the model.

7.Model

The random forest model is used for the observation of yields.

8.Validation data

The separate dataset from the training dataset used to tune the model's parameters and prevent overfitting.

9.Hyperparameter

Parameters like the maximum depth of the random tree is used for the prediction.

10. Epoch

One complete pass through all the crop yield used during the training process.

11.Loss function

Cross-entropy loss used to measure the difference between the actual yield and year by year yield.

12.Learning rate

The step size used for updating the yield of a crop during training.

13.Overfitting

The model predicts crop yield perfectly on training data but fails on test data, which indicates overfitting.

14.Underfitting

The model predicts crop yield poorly on both training and test data.

15.Regularization

Applying L2 regularization can prevent the model from overfitting the training data, improving generalization to new data.

16.Cross-validation

Performing k-fold cross-validation, where the dataset is split into k subsets, and the model is trained and tested k times, each time using a different subset as the test set and the remaining data as the training set.

17.Feature Engineering

Creating a new feature like fertilizer efficiency by combining the area and fertilizer usage to provide a more informative input to the model.

18.Dimensionality Reduction

Using PCA to reduce multiple features (different types of fertilizers) into a single component that captures the most variance in the data.

19.Bias

Systematic errors in models due to inadequate model.

20.Variance

High sensitivity to fluctuations in training data, causing the model to perform inconsistently.