MACHINE LEARNING
ASSIGNMENT 1

Plant ID	Age (weeks)	Temperature (°C)	Humidity (%)	Symptom 1	Symptom 2	Symptom 3	Diagnosis
001	8	25	70	Yellow leaves	Wilting	Spots	Nutrient Deficiency
002	12	28	60	Stunted growth	Leaf curling	Yellow edges	Virus Infection
003	10	22	80	White patches	Yellow leaves	Stunted growth	Fungal Infection
004	15	26	65	Brown spots	Leaf drop	Yellow leaves	Bacterial Infection
005	9	24	75	Yellowing veins	Brown edges	Drooping	Pest Infestation

Sample test data for Plant Disease Diagnosis :

The Terminologies used:

1. **Feature**

   - Plant's age, Temperature, Humidity, and Symptoms are the features of this dataset.

2. **Label**

   - The diagnosis of a specific plant disease.

3. **Prediction**

   - Predicting the disease of the plant based on symptoms.

4. **Outlier**

   - Unusually high or low temperature or humidity compared to the range of other plants.

5. **Test data**

   - The set of plant data used to test the diagnosis model.

6. **Training model**

   - The collection of plant records used to train the model to diagnose diseases.

7. **Model**

   - The decision tree model used to classify the disease of the plant by all the features.

8. **Validation data**

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

9. **Hyperparameter**

   - The maximum depth of the decision tree used for diagnosing the disease.

---

10. **Epoch**

   - One complete pass through the plant data during the training of the neural network for diagnosing the disease.

11. **Loss function**

   - Cross-entropy loss used to measure the difference between the predicted disease and actual disease.

12. **Learning rate**

   - The step size used for updating the weights of a neural network during training.

13. **Overfitting**

   - The model performs well on training data but poorly on test data because it learned even the noises and specifics.

14. **Underfitting**

   - The model is too simple and fails to diagnose complex problems, leading to poor performance.

15. **Regularization**

   - Adding a penalty to the model for large coefficients to prevent disease prediction models from overfitting.

16. **Cross-validation**

   - Splitting plant data into multiple subsets to test/train the model on different combinations to evaluate performance.

17. **Feature Engineering**

   - Creating a new feature from existing features, such as creating a new feature called 'stress index' by combining temperature and humidity features.

18. **Dimensionality Reduction**

   - Reducing the number of features in plant data using Principal Component Analysis (PCA).

19. **Bias**

   - Systematic errors in models due to an inadequate model.
   - High sensitivity to fluctuations in training data, causing the model to perform inconsistently.

20. **Variance**

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

BY :

SUGANTH A

22CSR210