**GENERATIVE AI & SYSTIMA NX JOIN TECH INTERNSHIP PROGRAM**

**ASSIGNMENT - I**

**VANJULA S  
22CSR229**

**DATASET TABLE:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **SAND**  **(%)** | **PH VALUE** | **ORGANIC MATTER**  **(%)** | **MOISTURE**  **CONTENT**  **(%)** | **CULTIVABLE**  **(Yes/No)** |
| 1 | 70 | 6.5 | 2.5 | 15 | Yes |
| 2 | 30 | 5.5 | 3.0 | 25 | No |
| 3 | 40 | 6.8 | 4.0 | 20 | Yes |
| 4 | 10 | 7.0 | 3.5 | 18 | Yes |
| 5 | 15 | 4.5 | 5.0 | 30 | No |

**TERMINOLOGIES:**

**1.Feature**

* **Definition**: Attributes or properties used to describe each data point in the dataset.
* **Examples**: Sand (%), pH, Organic Matter (%), Moisture Content (%) are features, for ID 1, the features are Sand (%) = 70, pH = 6.5, Organic Matter (%) = 2.5, Moisture Content (%) = 15.

**2.Label**

* **Definition**: The target variable or outcome that the model is trying to predict based on the features.
* **Example**: Cultivable (Label) indicating whether the soil is suitable for cultivation, for ID 3, the label is Yes, meaning the soil is suitable for cultivation.

**3.Prediction**

* **Definition**: It is the model’s attempt to predict the output label based on the input features.
* **Example**: If a model predicts Yes for ID 2 based on its features, it is predicting that the soil is suitable for cultivation.

**4.Outlier**

* **Definition:** A data point that significantly deviates from the other observations in the dataset. It doesn’t fit with the pattern of the other data.
* **Example:** Most of the soil samples have a pH around 6, but if one sample has a pH of 9, that sample could be an outlier.

**5.Test Data**

* **Definition:** A data subset used to evaluate the performance of the trained model. This data subset is not used during the training process.
* **Example:** In a split dataset, if IDs 1 to 4 are used for training, ID 5 could be used as test data to assess how well the model generalizes to new data.

**6.Training Data**

* **Definition**: The data subset used to train the ML model. It allows model to learn relationship between the features and label.
* **Example:** IDs 1 to 4 in the dataset are used to train the model to learn the relationship between various features and the Cultivable (Label).

**7.Model**

* **Definition:** An algorithm or method used to predict the label based on the features. Model is trained using data to make predictions.
* **Example:** A decision tree classifier or a logistic regression model can be used to predict soil cultivability based on these features.

**8.Validation Data**

* **Definition:** A subset of data used to tune model hyperparameters and assess performance during training.
* **Example:** If you use IDs 1, 2, and 3 for training, you might use ID 4 for validation to fine-tune the model before testing it on ID 5.

**9.Hyperparameter**

* **Definition:** Settings or configurations that are set before the model training begins and influence its performance.
* **Example:** In a decision tree model, the maximum depth of the tree is a hyperparameter.

**10.Epoch**

* **Definition:** One complete pass through the entire training dataset during the training process.
* **Example:** If a model is trained over 100 epochs, it means the model has processed the entire training data 100 times.

**11.Loss Function**

* **Definition**: A function used to measure the error between the model’s predictions and the actual values.
* **Example:** For binary classification, the cross-entropy loss function measures how well the predicted probabilities match the actual labels.

**12.Learning Rate**

* **Definition:** The rate at which the model’s weights are updated during training.
* **Example:** A learning rate of 0.01 means the model’s weights are adjusted by 0.01 times the gradient of the loss function.

**13.Overfitting**

* **Definition:** When a model learns the training data too well, including noise, leading to poor generalization on new data.
* **Example:** If a model performs very well on IDs 1 to 4 but poorly on ID 5, it may be overfitting the training data.

**14.Underfitting**

* **Definition:** When a model is too simple to capture the underlying patterns in the data, leading to poor performance on both training and test data.
* **Example:** A model that predicts all soil samples as “No” regardless of the actual features may be underfitting.

**15.Regularization**

* **Definition:** Techniques used to prevent overfitting by adding constraints or penalties to the model.
* **Example:** L1 (Lasso) or L2 (Ridge) regularization can be used in regression models to penalize large coefficients.

**16.Cross-Validation**

* **Definition:** A technique to assess model performance by splitting the data into multiple subsets and training/testing on different combinations.
* **Example:** In k-fold cross-validation, the dataset is divided into k subsets. The model is trained k times, each time using k-1 subsets for training and the remaining subset for testing.

**17.Feature Engineering**

* **Definition:** The process of creating new features or modifying existing ones to improve model performance.
* **Example:** Combining Sand (%) and Moisture Content (%) features to create a new feature, such as “Drainage Efficiency.”

**18.Dimensionality Reduction**

* **Definition:** Techniques which are used to reduce the number of features while retaining as much information as possible.
* **Example:** Using Principal Component Analysis (PCA) to reduce the feature set from Sand (%), pH, Organic Matter (%), and Moisture Content (%) to a smaller number of principal components.

**19.Bias**

* **Definition:** Systematic error introduced by the model, often due to oversimplification or assumptions.
* **Example:** A model with high bias might consistently predict all soils as “No” regardless of the actual features.

**20.Variance**

* **Definition:** The model’s sensitivity to fluctuations in the training data, leading to high variability in predictions.
* **Example:** A model with high variance might perform very well on the training data but poorly on new test data due to overfitting.