**Assignment 1**

**NAME: DEEPIKA P M**

**ROLLNO: 22CSR041**

**CLASS: III-A**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **AGE** | **FAMILY INCOME** | **STUDY HOURS** | **GRADE** | **SCHOLARSHIP AWARDED** |
| **1** | 18 | 30000 | 5 | A | No |
| **2** | 20 | 50000 | 3 | B | No |
| **3** | 22 | 80000 | 6 | A | Yes |
| **4** | 19 | 40000 | 4 | B | No |
| **5** | 21 | 60000 | 5 | B | Yes |
| **6** | 23 | 90000 | 7 | A | Yes |
| **7** | 24 | 100000 | 6 | A | Yes |
| **8** | 18 | 20000 | 2 | C | No |

 **Feature:**

Individual independent variables that act as inputs in the system.

Example: Age, Family Income, Study Hours.

 **Label:**

The target variable or the output we aim to predict.

Example: Scholarship Awarded (Yes/No).

 **Outlier:**

A data point that deviates significantly from other observations.

Example: ID=8 could be considered an outlier due to very low family income compared to others.

 **Test Data:**

Data used to evaluate the model's performance after training.

Example: Records of ID=6 and ID=7.

 **Training Data:**

Data used to train the model.

Example: Records from ID=1 to ID=5.

 **Model:**

Algorithm or program that makes predictions on new data based on what it learned during training.

Example: Logistic Regression, Random Forest.

 **Validation Data:**

Subset of data used to validate the model during training.

Example: Records of ID=3 and ID=4.

 **Hyperparameter:**

Parameters set before training to control the learning process.

Example: Learning rate, number of trees in Random Forest.

 **Epoch:**

One complete pass of the training dataset through the model.

Example: One pass through records of ID=1 to ID=5.

 **Loss Function:**

Measures how well the model's predictions match the actual target values.

Example: Cross-Entropy Loss.

 **Learning Rate:**

Controls the step size in optimization algorithms.

Example: Starting with a learning rate of 0.01 and adjusting based on performance.

 **Overfitting:**

Model performs well on training data but poorly on new, unseen data due to capturing noise rather than underlying patterns.

 **Underfitting:**

Model is too simple to capture the underlying patterns in the training data and fails to generalize to new data.

 **Regularization:**

Techniques to prevent overfitting by penalizing large coefficients.

Example: L1 Regularization (Lasso Regression).

 **Cross-validation:**

Technique to validate model performance by splitting data into multiple subsets for training and validation.

 **Feature Engineering:**

Process of creating new features from existing data to improve model performance. Example: Creating a combined score from Age and Study Hours.

 **Dimensionality Reduction:**

Reducing the number of input variables to focus on the most important features. Example: PCA (Principal Component Analysis).

 **Bias:**

Systematic error in the model due to incorrect assumptions in the learning process. Example: Selection bias in the dataset.

 **Variance:**

Model's sensitivity to small fluctuations in the training dataset.

Example: A complex model that varies greatly with small changes in the training data has high variance.