**Generative AI Consortium (Ltd)**

**AI/ML Internship: Assignment 1 (Understanding the basic ML terms)**

**Name: HARITHA E**

**Email:** [**harithaeswaran@gmail.com**](mailto:harithaeswaran@gmail.com)

**Sample Test Data: Engineering College Students’ Performance**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Student ID** | **Gender** | **Age** | **Major** | **Study Hours per Week** | **Internship (Yes/No)** | **GPA (out of 10)** |
| 1 | Female | 20 | |  | | --- | | Computer Science |  |  | | --- | |  | | 15 | Yes | 8.5 |
| 2 | Male | 21 | Mechanical Engineering | 12 | No | 7.8 |
| 3 | Female | 19 | Electrical Engineering | 18 | Yes | |  | | --- | | 9.2 |  |  | | --- | |  | |
| 4 | Male | 22 | Civil Engineering | 10 | No | 7.5 |
| 5 | Female | 20 | Chemical Engineering | 20 | Yes | 9.0 |
| 6 | Male | 21 | Computer Science | 14 | No | 7.4 |

FFfffff

**1. Feature:** Individual measurable properties or attributes used as input variables in the model. They help in predicting the target variable.

* **Example**: Age, Major, Study Hours per Week, Internship (Yes/No).

**2. Label:** The output variable that the model aims to predict. It is what you are trying to estimate or classify.

* **Example**: GPA (out of 10)

**3. Prediction:** The estimated value or outcome that the model provides based on the input features.

* **Example**: For a student with Age=20, internship and Study Hours per Week=15, the model might predict a GPA of 8.5.

**4. Outlier:** A data point that significantly differs from other observations in the dataset. It can be an anomaly or an extreme value.

* **Example**: A student with an unusually high GPA like 10.0.

**5. Test Data:** A subset of the dataset used to evaluate the performance of the model. This data is not used during training and helps assess how well the model generalizes to new data.

* **Example**: Records such as Student ID=2 and Student ID=5 might be used for testing the model’s predictions.

**6. Training data:** The subset of the dataset used to train the model. It helps the model learn the relationships between features and the label.

* **Example**: Records like Student ID=1, Student ID=3, and Student ID=4 might be used to train the model.

**7. Model:** An algorithm or mathematical framework that processes input features to make predictions or decisions.

* **Example**: A linear regression model or a decision tree used to predict student GPA based on features like Age, internship and Study Hours per Week.

**8. Validation Data:** A subset of the dataset used to tune the model's hyperparameters and validate its performance during training. It helps in fine-tuning the model and preventing overfitting.

* **Example**: Records such as Student ID=3 and Student ID=4

**9. Hyper parameter:** Parameters that are set before the model starts training and control the training process. They are not learned from the data but are configured beforehand.

* **Example**: The number of layers in a neural network or the depth of a decision tree.

**10. Epoch:** One complete pass through the entire training dataset during the training process. It represents a single iteration over all training data.

* **Example**: If the model trains on records from Student ID=1 to Student ID=4 for one epoch, it means it has seen each record once.

**11. Loss Function:** A function that measures the difference between the model’s predicted values and the actual target values. It helps in evaluating the model’s performance.

* **Example**: Mean Squared Error (MSE) measures the average squared difference between predicted GPA and actual GPA.

**12. Learning Rate:** A parameter that controls the step size during the optimization process. It determines how quickly or slowly the model adjusts its parameters.

* **Example**: A learning rate of 0.01 might be used to update the model’s parameters during training.

**13. Overfitting:** A condition where the model performs very well on training data but poorly on unseen data due to excessive complexity or memorization of training examples.

* **Example**: A model that perfectly predicts GPA for training data but fails to generalize to new student records.

**14. Underfitting:** When the 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 a constant GPA for all students regardless of their features.

**15. Regularization**: Techniques used to prevent overfitting by adding penalties to the loss function, discouraging overly complex models.

* **Example**: L2 Regularization, which adds a penalty proportional to the square of the model’s coefficients.

**16. Cross-validation:** A technique to evaluate the model’s performance by splitting the dataset into multiple parts, training and testing the model on different subsets.

* **Example**: 5-fold cross-validation, where the dataset is divided into five parts and the model is trained and tested on each part in turn.

**17. Feature Engineering:** The process of creating new features or modifying existing ones to improve the model’s performance and provide better insights.

* **Example**: Creating a Study Intensity feature by combining Study Hours per Week with Major.

**18. Dimensional Reduction:** Techniques used to reduce the number of features in the dataset while retaining important information.

* **Example**: Principal Component Analysis (PCA) used to simplify the dataset by reducing the number of features.

**19. Bias**: Systematic errors introduced by incorrect assumptions or limitations in the model.

* **Example**: Assuming a linear relationship between Study Hours per Week and GPA when the relationship is actually more complex.

**20. Variance**: The variability in the model’s predictions due to changes in the training data. High variance indicates that the model is overly sensitive to fluctuations in the data.

* **Example**: A model that produces significantly different GPA predictions when trained on slightly different subsets of data.