**Generative AI Consortium (Ltd)**

**AI/ML Internship: Assignment 1 (Simple Machine Learning Problem) Name: KEERTHANA S**

**Email:** [**bhanukeerthi2002@gmail.com**](mailto:bhanukeerthi2002@gmail.com)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Student ID** | **Hours studied** | **Attendance**  **%** | **Assignments Completed** | **Previous GPA** | **Grade** |
| 1 | 30 | 90 | 10 | 3.5 | A |
| 2 | 20 | 85 | 8 | 3.0 | B |
| 3 | 25 | 80 | 9 | 3.2 | B+ |
| 4 | 10 | 75 | 6 | 2.8 | C |
| 5 | 35 | 95 | 10 | 3.8 | A+ |
| 6 | 40 | 100 | 12 | 3.9 | A+ |
| 7 | 15 | 70 | 5 | 2.5 | D |

# Feature:

**Feature**: Individual measurable properties or characteristics used as inputs to the model.

**Example**: Hours Studied, Attendance , Assignments Completed, Previous GPA.

# Label:

**Label** : The output variable that the model aims to predict.

**Example**: Grade.

# Prediction:

**Prediction**: The output produced by a machine learning model when it is applied to a set of input features.

**Example**: Predicting the grade of a student.

# Outlier:

**Outlier**: A data point that deviates significantly from the rest of the data.

**Example**: A student with an extremely high or low grade compared to others.

# Test Data:

**Test Data** : A subset of the dataset used to provide an unbiased evaluation of a model fit on the training dataset.

**Example**: student grade

# Training Data:

**Training Data** : The dataset used to train a machine learning model.

**Example**: The sample data provided.

# Model:

**Model** : An algorithm or mathematical representation that maps input features to the desired output.

**Example**: A regression model predicting student grades.

# Validation Data:

**Validation Data** : A subset of the dataset used to tune the model's hyperparameters.

**Example**: A portion of the data used to validate the model's performance.

# Hyperparameter

**Hyperparameter :** Parameters that are set before training a model and controlling the learning process.

**Example**: Learning rate, number of epochs.

# Epoch

**Epoch** : One complete pass through the entire training dataset.

**Example**: If the model is trained over 100 epochs, the training dataset is processed 100 times.

# Loss Function

**Loss Function:** Quantifies the difference between predicted outputs of a machine learning algorithm and actual target values.

**Example**:Mean Squared Error (MSE) in regression problems.

# Learning Rate

**Learning Rate:** Tuning parameter in an optimization algorithm that determines the step size at each iteration while moving towards a minimum of a loss function.

**Example:** A learning rate of 0.01.

# Overfitting

**Overfitting:** A behaviour that occurs when the learning model gives accurate predictions for training data but not for new data.

**Example**: A model that has very low error on training data but high error on test data.

# Underfitting

**Underfitting:** When a model is too simple and has not learned the patterns in the training data well and is unable to generalize well on the new data.

**Example**: A linear model applied to highly non-linear data.

# Regularization

**Regularization**: Set of methods to reduce overfitting.

**Example**: Lasso (L1) and Ridge (L2)

# Cross-Validation

**Cross-Validation:** Technique of resampling different portions of training data for validation on different iterations.

**Example**: k-fold cross-validation.

# Feature Engineering

**Feature Engineering:** Technique that leverages data to create new variables that aren’t in the training set.

**Example**: Creating a feature that represents the total study time from hours studied per day.

# Dimensionality Reduction:

**Dimensional Reduction:** Method of reducing variables in a training dataset used to develop machine learning models.

**Example**: Principal component analysis(PCA).

# Bias

**Bias**: Systematic error that occurs in the model itself due to incorrect assumptions on the machine learning process.

**Example**: Sample Bias.

# Variance:

**Variance**: Changes in the model when using different portions of the training dataset.

**Example**:A model that predicts widely varying grades for similar students due to noise in the training data.