|  |
| --- |
| **Generative AI Consortium (Ltd)**  **AI/ML Internship: Assignment 1 (Simple Machine Learning Problem)**  **Name: ADITHYAN M** |
| **Email:** [**mailto:adithyanlachu@gmail.com**](mailto:adithyanlachu@gmail.com)   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **ID** | **Age** | **Hours Studied** | **Attendance (%)** | **Assignments Submitted** | **Participation** | **Test Score (%)** | | 1 | 18 | 10 | 95 | 8 | High | 88 | | 2 | 19 | 8 | 85 | 7 | Medium | 75 | | 3 | 17 | 12 | 100 | 10 | High | 92 | | 4 | 20 | 6 | 70 | 6 | Low | 65 | | 5 | 18 | 9 | 80 | 8 | Medium | 78 | | 6 | 21 | 5 | 60 | 5 | Low | 55 | |

FFfffff

**Feature:** These are the individual measurable properties or characteristics used as inputs to the model. In this dataset, the features include Age, Hours Studied, Attendance (%), Assignments Submitted, and Participation.

**Label:** The output variable that the model aims to predict. Here, the label is Test Score (%).

**Prediction:** The estimated test score generated by the model based on the input features.

**Outlier:** A data point that deviates significantly from other data points. For example, if a student had a Test Score of 100% despite very low Hours Studied and Attendance (%), it might be considered an outlier.

**Test Data:** A subset of this dataset that is used to evaluate the performance of the model. It is separate from the training data.

**Training Data:** The portion of the dataset used to train the model. For instance, IDs 1-4 could be used as training data, and IDs 5-6 as test data.

**Model:** A representation that learns the relationship between features and labels from the training data. In this case, the model would learn to predict test scores based on features like age, hours studied, etc.

**Validation Data:** A subset of data used to tune the model's hyperparameters. This is separate from both the training and test data.

**Hyperparameter:** Configurable parameters that are set before the learning process begins, like learning rate or the number of trees in a random forest.

**Epoch:** A single pass through the entire training dataset during the training process.

**Loss Function:** A function that measures the difference between the predicted and actual test scores, such as Mean Squared Error (MSE).

**Learning Rate:** A hyperparameter that controls how much to change the model in response to the estimated error.

**Overfitting:** When the model learns the training data too well, including noise and outliers, leading to poor generalization to new data.

**Underfitting:** When the model is too simple to capture the underlying pattern in the data, resulting in poor performance.

**Regularization:** Techniques used to prevent overfitting by penalizing model complexity.

**Cross-Validation:** A technique for evaluating the model's performance by splitting the data into multiple subsets and training/testing on different combinations.

**Feature Engineering:** The process of selecting, modifying, or creating new features to improve the performance of the model.

**Dimensionality Reduction:** Techniques used to reduce the number of input features in a dataset.

**Bias:** An error introduced by approximating a real-world problem by a simpler model. High bias can cause underfitting.

**Variance:** The model's sensitivity to fluctuations in the training data. High variance can lead to overfitting.