|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Generative AI Consortium (Ltd)**   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | |  | | --- | | **ML Internship: Assignment 1**  **Name: VARUNIKAA K** | | **Email:** varunikaak.22cse@kongu.edu | | | | | | | **STUDENT ID** | **STUDY HOURS** | **CLASS ATTENDENCE(%)** | **PREVIOUS TEST SCORE** | **FINAL GRADE** | | 1 | 15 | 85 | 80 | B | | 2 | 10 | 75 | 70 | C | | 3 | 20 | 90 | 95 | A | | 4 | 5 | 60 | 50 | D | | 5 | 25 | 95 | 85 | A | |  |  |  |  |

**TERMINOLOGIES**

**Feature:** Individual measurable properties or characteristics of the data

Example: The features are **Study Hours**, **Class Attendance (%)**, and **Previous Test Scores**.

**Label:** The target variable that we want to predict.

Example: The label is the **Final Grade**.

**Prediction:** The output of the model based on the input features.

Example: The model might predict a final grade of **B** for a student with 15 study hours, 85% attendance, and 80 on previous tests.

**Outlier:** A data point that differs significantly from other observations.

Example: If there was a student with 50 study hours, 100% attendance, and a previous test score of 100 but received a final grade of **F**, that would be an outlier.

**Test Data:** A subset of the dataset used to evaluate the model's performance. It is not used during the training of the model.

Example: Students 4 and 5 could be used as test data.

**Training Data:** A subset of the dataset used to train the model.

Example: Students 1, 2, and 3 could be used as training data.

**Model:** A mathematical representation that makes predictions based on input data. In this case, it might be a logistic regression model predicting final grades.

**Validation Data:** A subset of the dataset used to tune the model parameters and prevent overfitting.

**Hyperparameter:** Parameters set before the training process begins, such as learning rate or regularization strength.

**Epoch:** One complete pass through the entire training dataset. For instance, if the model is trained on Students 1, 2, and 3, one epoch means that all three students have been processed once.

**Loss Function:** A function that measures how well the model's predictions match the actual labels

**Learning Rate:** A hyperparameter that controls how much the model's parameters are adjusted during training.

**Overfitting:** When the model performs well on the training data but poorly on new, unseen data because it has learned the noise in the training data.

Example: If the model predicts the exact final grades of the training students but fails to generalize to new students, it is overfitting.

**Underfitting:** When the model is too simple to capture the underlying pattern in the data, performing poorly on both training and test data.

Example: If a model assumes grades depend only on study hours and ignores attendance and previous scores, it might underfit.

**Regularization:** Techniques used to prevent overfitting by adding a penalty for more complex models.

**Cross-Validation:** A technique for evaluating the model's performance by splitting the data into multiple training and validation sets.

**Feature Engineering:** Creating new features or modifying existing ones to improve model performance.

Example: Creating a new feature like "Study Hours per Week" might provide more insight for the model.

**Dimensionality Reduction:** Techniques to reduce the number of features while preserving important information

**Bias:** Error introduced by approximating a real-world problem with a simplified model. High bias can cause underfitting. For instance, assuming final grades depend only on study hours might introduce bias.

**Variance:** Error introduced by the model's sensitivity to small fluctuations in the training data. High variance can cause overfitting. For instance, if the model learns the noise in the training data, it will have high variance.