|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Generative AI Consortium (Ltd)**  **AI/ML Internship: Assignment 1 (Simple Machine Learning Problem) Assignment)**  **Name: DHARSHINI M K** | | **Email:** [**mkdharshini24@gmail.com**](mailto:mkdharshini24@gmail.com) | | | | | | |  |
| **ID** | **Age** | **Salary** | **Purchased** | **Income Level** | **Is Outlier** |
| 1 | 25 | 50000 | No | Medium | No |
| 2 | 45 | 100000 | Yes | High | No |
| 3 | 35 | 75000 | No | Medium | No |
| 4 | 50 | 120000 | Yes | High | No |
| 5 | 23 | 40000 | No | Low | No |
| 6 | 40 | 90000 | Yes | High | No |
| 7 | 120 | 5000000 | No | Outlier | Yes |

FFfffff

**Feature:** Individual independent variables that act like an input in your system.

**Example**: Age, Salary, Income level.

**Label:** Identification of raw data.

**Example**: Purchased.

**Prediction:** Project a probable dataset that relates back to original data.

**Example**: For a new record in the dataset with Age=30 and Salary=60000, the model might predict No.

**Outlier:** Data that is unique/different from other data.

**Example**: id=7 where the outlier=yes.

**Test Data:** Ensure that the model works for the given testing data.

**Example**: Records of id=6 and id=7.

**Training data:** Data that is used to train the model.

**Example**: Records from id=1 to id=5.

**Model:** Program that can make decisions from previously unseen datasets.

**Example**: Ensemble trees, Decision tree.

**Validation Data:** Uses a sample of data that is with-held from training.

**Example**: Records of id=3 and id=4.

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

**Example**: The topology and size of a neural network.

**Epoch:** Each time a dataset passes through an algorithm, it is said to have completed one epoch. Therefore it refers to the one complete passing of training data through algorithm.

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

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

**Example**: Mean Square Error, Mean Absolute Error.

**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**: Starting with a learning rate of 0.1 and reduce it by a factor of 0.5 every 10 epochs.

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

**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.

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

**Example**: L2 Regularization.

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

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

**Example**: Creating income level by binning salary into categories like low, medium and high.

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

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

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

**Example**: Sample Bias.

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

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