|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | **Generative AI Consortium (Ltd)**  **AI/ML Internship: Assignment 1 (Simple Machine Learning Problem) Assignment)**  **Name: HARINEE N J** | | **Email: njharinee1111@gmail.com** | |  |

**Employee Performance**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Employee ID** | |  | | --- | |  |  |  | | --- | | **Age** | | **Department** | **Education Level** | **Years at Company** | **Monthly Hours** | **Performance Rating** |
| 1 | 25 | Sales | Bachelor's | 2 | 160 | 3 |
| 2 | 35 | IT | |  | | --- | |  |  |  | | --- | | Master's | | 5 | 180 | 4 |
| 3 | 24 | HR | Bachelor's | 3 | 150 | 2 |
| 4 | 42 | IT | |  | | --- | |  |  |  | | --- | | Master's | | 10 | 200 | 5 |
| 5 | 30 | Sales | Bachelor's | 4 | 170 | 3 |
| 6 | 38 | HR | |  | | --- | |  |  |  | | --- | | Master's | | 7 | 160 | 4 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

FFfffff

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

**Example:** In this dataset, features are Age, Department, Education Level, Years at Company, and Monthly Hours.

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

**Example:** In this dataset, the label is Performance Rating.

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

**Example:** For a new record in the dataset with Age=30 and Department=Sales, the model might predict a performance rating of 3.

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

**Example:** If there was an employee with a performance rating of 1 while the rest are between 2 and 5, that would be an outlier.

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

**Example:** Records of Employee ID=5 and Employee ID=6.

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

**Example:** Records from Employee ID=1 to Employee ID=4.

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

**Example:** A decision tree model predicting performance ratings based on employee features.

**Validation Data:** Uses a sample of data that is withheld from training.

**Example:** Records of Employee ID=3 and Employee ID=4.

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

**Example:** The number of trees in a random forest, or the learning rate in a neural network.

**Epoch:** Each time a dataset passes through an algorithm, it is said to have completed one epoch.

**Example:** One complete pass through records from Employee ID=1 to Employee ID=4 in the training data.

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

**Example:** Mean Squared Error (MSE) used to measure the difference between predicted and actual performance ratings.

**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 0.01 and reducing it by a factor of 0.5 every 10 epochs.

**Overfitting:** Occurs when the learning model gives accurate predictions for training data but not for new data.

**Example:** If our model predicts performance ratings perfectly for the training data but poorly for the test 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.

**Example:** If our model predicts performance ratings poorly for both training and test data.

**Regularization:** Set of methods to reduce over fitting.

**Example:** L2 regularization applied to our model to penalize large coefficients.

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

**Example:** Using 5-fold cross-validation to validate the performance of our model.

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

**Example:** Creating a new feature like Years per Department by dividing Years at Company by the number of departments an employee has worked in.

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

**Example:** Using Principal Component Analysis (PCA) to reduce the features from 5 to 2 while retaining most of the variance.

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

**Example:** Sample bias if our training data only includes employees from one department.

**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.