|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Generative AI Consortium (Ltd)**  **AI/ML Internship: Assignment 1 (Simple Machine Learning Problem) Assignment)**  **Name: DINESH KARTHIK RAJAN D** | | **Email:** [**mailto:dineshkarthikrajand@gmail.com**](mailto:dineshkarthikrajand@gmail.com) | |  | | | | | | |  |  |  |
| **ID** | **Age** | **Salary** | **Purchased** | **Income Level** | **Is Outlier** | **Gender** | **Education Level** |
| 1 | 28 | 56000 | No | Medium | No | Male | Bachelor |
| 2 | 42 | 80000 | Yes | High | No | Female | Master |
| 3 | 35 | 70000 | Yes | Medium | No | Male | Bachelor |
| 4 | 50 | 150000 | Yes | High | No | Female | PhD |
| 5 | 22 | 30000 | No | Low | No | Male | High School |
| 6 | 40 | 85000 | Yes | High | No | Female | Master |
| 7 | 70 | 200000 | Yes | High | No | Male | Bachelor |

FFfffff

1. **Feature**: Individual independent variables that act as input in your system.
   * **Example**: Age, Salary, Income Level, Gender, Education Level.
2. **Label**: Identification of raw data.
   * **Example**: Purchased.
3. **Prediction**: Project a probable dataset that relates back to the original data.
   * **Example**: For a new record with Age=30 and Salary=60000, the model might predict No.
4. **Outlier**: Data that is unique/different from other data.
   * **Example**: ID=13 in the previous dataset where the Outlier=Yes.
5. **Test Data**: Ensure that the model works for the given testing data.
   * **Example**: Records of ID=6 and ID=7.
6. **Training Data**: Data that is used to train the model.
   * **Example**: Records from ID=1 to ID=5.
7. **Model**: Program that can make decisions from previously unseen datasets.
   * **Example**: Decision Tree, Logistic Regression.
8. **Validation Data**: Uses a sample of data that is withheld from training.
   * **Example**: Records of ID=3 and ID=4.
9. **Hyperparameter**: Parameters that are set before training a model and control the learning process.
   * **Example**: The depth of the decision tree or the learning rate for gradient boosting.
10. **Epoch**: Each time a dataset passes through an algorithm, it is said to have completed one epoch.
    * **Example**: One pass through records of ID=1 to ID=5.
11. **Loss Function**: Quantifies the difference between predicted outputs of a machine learning algorithm and actual target values.
    * **Example**: Mean Square Error, Cross-Entropy Loss.
12. **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 reduce it by a factor of 0.1 every 20 epochs.
13. **Overfitting**: A behavior that occurs when the learning model gives accurate predictions for training data but not for new data.
    * **Example**: A model that performs very well on the training data but poorly on the test data.
14. **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 model that performs poorly on both the training and test data, indicating it has not captured the underlying patterns in the data.
15. **Regularization**: Set of methods to reduce overfitting.
    * **Example**: L2 Regularization (Ridge Regression).
16. **Cross-validation**: Technique of resampling different portions of training data for validation on different iterations.
    * **Example**: Using k-fold cross-validation where k=5.
17. **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.
18. **Dimensional Reduction**: Method of reducing variables in a training dataset used to develop machine learning models.
    * **Example**: Principal Component Analysis (PCA).
19. **Bias**: Systematic error that occurs in the model itself due to incorrect assumptions on the machine learning process.
    * **Example**: Sample Bias.
20. **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.