|  |
| --- |
| **Generative AI Consortium (Ltd)**  **AI/ML Internship: Assignment 1 (Simple Machine Learning Problem) Assignment)**  **Name: ANAND S** |
| **Email:** [**mailto:anandsaravanan38@gmail.com**](mailto:anandsaravanan38@gmail.com) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Age** | **Salary** | **Purchased** | **Income Level** | **Is Outlier** |
| **1** | **28** | **55000** | **No** | **Medium** | **No** |
| **2** | **42** | **95000** | **Yes** | **High** | **No** |
| **3** | **37** | **80000** | **No** | **Medium** | **No** |
| **4** | **53** | **130000** | **Yes** | **High** | **No** |
| **5** | **22** | **42000** | **No** | **Low** | **No** |
| **6** | **47** | **110000** | **Yes** | **High** | **No** |
| **7** | **115** | **4500000** | **No** | **Outlier** | **Yes** |

**Feature**: Independent variables used as inputs. Examples: Age, Salary, Income Level.

**Label**: Identification of raw data. Example: Purchased.

**Prediction**: Estimating probable outcomes based on input data. Example: For a new record with Age=32 and Salary=65000, the model might predict No.

**Outlier**: Unique or significantly different data. Example: ID=7 where Is Outlier=Yes.

**Test Data**: Data used to ensure the model's performance. Example: Records of ID=6 and ID=7.

**Training Data**: Data used to train the model. Example: Records from ID=1 to ID=5.

**Model**: A program that makes decisions from new datasets. Examples: Decision Tree, Random Forest.

**Validation Data**: A subset of data held back from training. Example: Records of ID=3 and ID=4.

**Hyperparameter**: Pre-set parameters controlling the learning process. Example: Number of trees in a Random Forest.

**Epoch**: One complete pass of the training data through the algorithm. Example: One pass through records of ID=1 to ID=5.

**Loss Function**: Measures the difference between predicted outputs and actual values. Examples: Mean Squared Error, Cross-Entropy Loss.

**Learning Rate**: A parameter that controls the step size during optimization. Example: Starting with a learning rate of 0.01 and reducing it gradually.

**Overfitting**: When the model fits the training data too well but performs poorly on new data.

**Underfitting**: When the model is too simple to capture patterns in the data, leading to poor performance on both training and new data.

**Regularization**: Techniques to reduce overfitting. Example: L1 Regularization.

**Cross-Validation**: Resampling method to validate the model on different training data subsets.

**Feature Engineering**: Creating new features from existing data. Example: Creating a new feature called "Age Group" by categorizing ages.

**Dimensionality Reduction**: Reducing the number of features. Example: Using PCA to reduce dimensions.

**Bias**: Systematic error in the model due to incorrect assumptions. Example: Measurement Bias.

**Variance**: Model's sensitivity to changes in the training data. Example: A high-variance model changes significantly with different training data subsets.