|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Generative AI Consortium (Ltd)**  **AI/ML Internship: Assignment 1 (Simple Machine Learning Problem) Assignment)**  **Name: ELANKUMARAN R** | | **Email:** [**mailto:elankumaran2103@gmail.com**](mailto:elankumaran2103@gmail.com) | | | | | | |  |
| **ID** | **Age** | **Salary** | **Purchased** | **Income Level** | **Is Outlier** |
| 1 | 22 | 48000 | No | Low | No |
| 2 | 28 | 60000 | Yes | Medium | No |
| 3 | 35 | 75000 | Yes | High | No |
| 4 | 40 | 90000 | Yes | High | No |
| 5 | 45 | 110000 | Yes | High | No |
| 6 | 50 | 130000 | Yes | High | No |
| 7 | 55 | 150000 | Yes | High | No |
| 8 | 60 | 180000 | Yes | High | No |
| 9 | 65 | 200000 | Yes | High | No |
| 10 | 70 | 220000 | Yes | High | No |

FFfffff

 **Feature:**  
Individual independent variables that act as inputs in the system or model.  
**Examples:** Age, Salary, Income Level.

 **Label:**  
The variable being predicted or classified.  
**Example:** Purchased (whether a person purchased or not).

 **Prediction:**  
Projecting a probable outcome based on input variables.  
**Example:** Predicting "Yes" for a new record with Age=25 and Salary=70000.

 **Outlier:**  
A data point significantly different from others.  
**Example:** There are no outliers in this dataset.

 **Test Data:**  
Data used to evaluate the model after training.  
**Example:** Records of ID 7 and ID 10 could be used as test data.

 **Training Data:**  
Data used to train the machine learning model.  
**Example:** Records from ID 1 to ID 6.

 **Model:**  
A program that makes decisions from new, unseen datasets.  
**Example:** Decision tree, Support Vector Machine.

 **Validation Data:**  
Data used to fine-tune model parameters and assess performance during training.  
**Example:** Records of ID 3 and ID 4 could be used for validation.

 **Hyperparameter:**  
Parameters set before training to control the learning process.  
**Example:** Learning rate, number of hidden layers.

 **Epoch:**  
One complete pass of the training dataset through a machine learning algorithm.  
**Example:** One epoch completes when all records from ID 1 to ID 10 have been used for training.

 **Loss Function:**  
Measures the difference between predicted and actual values.  
**Example:** Mean Squared Error, Cross-entropy loss.

 **Learning Rate:**  
Controls how much to change the model in response to the estimated error each time the model weights are updated.  
**Example:** Starting with a learning rate of 0.1 and adjusting based on validation performance.

 **Overfitting:**  
A behaviour where the model learns too much from the training data and performs poorly on new data.  
**Example:** If a complex model (like a deep neural network) memorizes the training data but fails to generalize to unseen data.

 **Underfitting:**  
When a model is too simple to capture the underlying patterns in the training data and performs poorly overall.  
**Example:** A linear model trying to fit a complex non-linear relationship in the data.

 **Regularization:**  
Techniques to prevent overfitting by penalizing large coefficients in the model.  
**Example:** L2 regularization (Ridge regression), L1 regularization (Lasso regression).

 **Cross-validation:**  
Technique to assess model performance by splitting data into multiple subsets for training and validation.  
**Example:** Using k-fold cross-validation to train the model on different subsets and validate it on others.

 **Feature Engineering:**  
Creating new features from existing data to improve model performance.  
**Example:** Creating a new feature "Age Group" from "Age" to categorize individuals into age ranges.

 **Dimensionality Reduction:**  
Techniques to reduce the number of input variables in a dataset.  
**Example:** Principal Component Analysis (PCA) to transform high-dimensional data into a smaller set of principal components.

 **Bias:**  
Systematic error that affects the accuracy of a model.  
**Example:** If the dataset is biased towards certain demographics, the model may not generalize well to a diverse population.

 **Variance:**  
Sensitivity of a model to changes in the training data.  
**Example:** A model with high variance may perform well on training data but poorly on test data due to its sensitivity to small fluctuations in the training set.