**Generative AI Consortium (Ltd)**

**AI/ML Internship: Assignment 1 (Simple Machine Learning Problem) Assignment)**

**Name: AKASHINE Y**

**Email:** [**mailto:yuvarajakashine@gmail.com**](mailto:yuvarajakashine@gmail.com)

| **ID** | **Age** | **Credit Score** | **Loan Amount** | **Approved** | **Income Level** | **Is Outlier** |
| --- | --- | --- | --- | --- | --- | --- |
| **1** | **30** | **650** | **10000** | **Yes** | **Medium** | **No** |
| **2** | **45** | **700** | **20000** | **Yes** | **High** | **No** |
| **3** | **25** | **600** | **5000** | **No** | **Low** | **No** |
| **4** | **50** | **720** | **15000** | **Yes** | **High** | **No** |
| **5** | **35** | **580** | **7000** | **No** | **Medium** | **No** |
| **6** | **29** | **690** | **12000** | **Yes** | **Medium** | **No** |
| **7** | **40** | **300** | **30000** | **No** | **Low** | **Yes** |

1. Feature: Individual independent variables that act like inputs in your system.
   * Example: Age, Credit Score, Loan Amount, Income Level.
2. Label: The target variable that the model is trying to predict.
   * Example: Approved.
3. Prediction: The model's output for a given input.
   * Example: For a new record with Age=28, Credit Score=680, and Loan Amount=8000, the model might predict Yes.
4. Outlier: A data point that is significantly different from other data points.
   * Example: Record with ID=7 where Is Outlier=Yes.
5. Test Data: Data used to evaluate the performance of the trained model.
   * Example: Records with ID=6 and ID=7.
6. Training Data: Data used to train the model.
   * Example: Records with ID=1 to ID=5.
7. Model: A mathematical representation learned from training data to make predictions.
   * Example: Logistic Regression, Decision Tree.
8. Validation Data: Data used to tune hyperparameters and prevent overfitting during training.
   * Example: Records with ID=3 and ID=4.
9. Hyperparameter: Settings used to control the learning process before training begins.
   * Example: Regularization strength in logistic regression, max depth of a decision tree.
10. Epoch: One complete pass through the entire training dataset.
    * Example: One pass through records with ID=1 to ID=5.
11. Loss Function: A measure of how well the model's predictions match the actual data.
    * Example: Binary Cross-Entropy Loss for classification tasks.
12. Learning Rate: A hyperparameter that controls how much to change the model in response to the error each time the model weights are updated.
    * Example: Starting with a learning rate of 0.01 and reducing it by a factor of 0.1 every 20 epochs.
13. Overfitting: When a model performs well on training data but poorly on unseen data.
    * Example: A model that has 100% accuracy on training data but low accuracy on test data.
14. Underfitting: When a model is too simple to capture the underlying patterns in the data.
    * Example: A linear model that fails to capture the relationship between credit score and loan approval.
15. Regularization: Techniques used to prevent overfitting by adding a penalty to the model.
    * Example: L1 and L2 regularization in linear models.
16. Cross-Validation: A technique for evaluating how a model generalizes to an independent dataset.
    * Example: 5-fold cross-validation where the dataset is divided into 5 parts and the model is trained and validated 5 times.
17. Feature Engineering: Creating new features from existing data to improve model performance.
    * Example: Creating a new feature "Debt-to-Income Ratio" from "Loan Amount" and "Income Level".
18. Dimensionality Reduction: Reducing the number of features in the dataset.
    * Example: Using Principal Component Analysis (PCA) to reduce features.
19. Bias: The error introduced by approximating a real-world problem with a simplified model.
    * Example: A model that assumes a linear relationship between credit score and loan approval might have high bias.
20. Variance: The error introduced by the model's sensitivity to small fluctuations in the training set.
    * Example: A highly complex model might change significantly with small changes in training data, indicating high variance.

Let's use the dataset to explain each term in context:

1. Features: Age, Credit Score, Loan Amount, Income Level.
2. Label: Approved.
3. Prediction: The model might predict loan approval based on input features.
4. Outlier: The record with ID=7.
5. Test Data: Records with ID=6 and ID=7.
6. Training Data: Records with ID=1 to ID=5.
7. Model: A logistic regression model could be used to predict loan approval.
8. Validation Data: Records with ID=3 and ID=4 used during training for validation.
9. Hyperparameter: Learning rate set to 0.01, regularization strength.
10. Epoch: Training the model through one complete pass of the training data.
11. Loss Function: Using binary cross-entropy loss to measure the difference between predicted and actual loan approvals.
12. Learning Rate: 0.01, controlling the step size during optimization.
13. Overfitting: If the model performs perfectly on training data but poorly on test data.
14. Underfitting: If the model fails to capture the relationship between features and loan approval.
15. Regularization: Adding L2 regularization to the logistic regression model to prevent overfitting.
16. Cross-Validation: Using 5-fold cross-validation to ensure the model generalizes well.
17. Feature Engineering: Creating a new feature "Debt-to-Income Ratio".
18. Dimensionality Reduction: Applying PCA to reduce the number of features.
19. Bias: If the model consistently predicts wrong outcomes due to incorrect assumptions.
20. Variance: If the model's predictions vary significantly with small changes in training data.