**Generative AI Consortium (Ltd)**

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

**Name: ASHILIN.B.S**

**Email:** [**ashilin04@gmail.com**](mailto:ashilin04@gmail.com)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Employee ID** | **Years of Experience** | **Education Level** | **Age** | **Department** | **Salary (USD)** |
| 1 | 5 | Bachelor's | 30 | HR | 50000 |
| 2 | 10 | Master's | 40 | IT | 80000 |
| 3 | 3 | Bachelor's | 25 | Marketing | 45000 |
| 4 | 15 | PhD | 50 | IT | 120000 |
| 5 | 8 | Master's | 35 | HR | 70000 |

1. **Feature**: Individual measurable properties or characteristics used as input to the model.
   * **Example**: Years of Experience, Education Level, Age, Department.
2. **Label**: The output variable that the model aims to predict.
   * **Example**: Salary.
3. **Prediction**: The value that the model estimates for the label based on the features.
   * **Example**: Predicted salaries based on the given features.
4. **Outlier**: A data point that deviates significantly from the rest of the data.
   * **Example**: If there was an employee with a salary of $250,000 in this dataset.
5. **Test Data**: A subset of the data used to assess the performance of a trained model.
   * **Example**: If we use records 4 and 5 to evaluate the model.
6. **Training Data**: The subset of the dataset used to train the model.
   * **Example**: Records 1, 2, and 3 can be used to train the model.
7. **Model**: An algorithm or mathematical representation that makes predictions based on the features.
   * **Example**: A regression model predicting employee salaries.
8. **Validation Data**: A subset of the data used to tune the model's hyperparameters.
   * **Example**: If we use record 4 to validate the model's performance during training.
9. **Hyperparameter**: Configuration settings used to tune the model's learning process.
   * **Example**: Learning rate, number of epochs.
10. **Epoch**: One complete pass through the entire training dataset.
    * **Example**: Training the model on records 1, 2, and 3 once.
11. **Loss Function**: A method to measure how well the model's predictions match the actual data.
    * **Example**: Mean Squared Error (MSE) between predicted and actual salaries.
12. **Learning Rate**: A hyperparameter that controls how much the model adjusts the weights with each update.
    * **Example**: A learning rate of 0.01.
13. **Overﬁtting**: When the model learns the training data too well, including its noise, and performs poorly on new data.
    * **Example**: If our model predicts the training data perfectly but fails on test data.
14. **Underﬁtting:** When the model is too simple to capture the underlying patterns in the data and performs poorly on both training and test data.

# Example:

* + - Using a very simple model (e.g., predicting salary based solely on age) might fail to capture other important factors like education level or experience, leading to poor predictions for all records.

1. **Regularization:** Techniques used to prevent overfitting by adding a penalty to the model’s complexity.

# Example:

* + - **Lasso (L1 regularization)** adds a penalty equal to the absolute value of the coefficients.
    - **Ridge (L2 regularization)** adds a penalty equal to the square of the coefficients

1. **Cross-Validation:** A technique to assess how the model generalizes to an independent dataset by splitting the data into multiple subsets, training the model on some subsets, and validating it on others.

# Example:

o Using 5-fold cross-validation, the dataset is split into 5 parts. The model is trained on 4 parts and tested on the remaining part, and this process is repeated 5 times with each part used as a test set once

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

**Example:** Creating a Salary Category feature by binning Salary into categories like Low (<$60,000), Medium ($60,000-$80,000), and High (>$80,000).

1. **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 Years of Experience, Age, and Education Level into principal components that capture the most variance in Salary.

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

# Example

**Sample Bias:** A model trained only on data from a specific geographic region may not generalize well to other regions.

1. **Variance:** Changes in the model when using diﬀerent portions of the training dataset, indicating sensitivity to data ﬂuctuations.

**Example of High Variance:** A highly complex regression model that fits the training data (e.g., records 1-3) very closely, but predicts wildly diﬀerent salaries for test data (e.g., records 4-5) due to its sensitivity to the training data.