### 

| **Generative AI Consortium (Ltd)****AI/ML Internship: Assignment 1 (Simple Machine Learning Problem)****Name: ARUNA A** **Email: arunaananthagiri04@gmail.com** |
| --- |

### 

| **Age** | **BMI** | **Blood Pressure** | **Glucose Level** | **Insulin Level** | **Has Diabetes** |
| --- | --- | --- | --- | --- | --- |
| 45 | 25 | 120 | 85 | 130 | Yes |
| 50 | 30 | 140 | 90 | 150 | Yes |
| 30 | 20 | 110 | 70 | 100 | No |
| 40 | 27 | 130 | 80 | 140 | Yes |
| 35 | 22 | 115 | 75 | 110 | No |
| 55 | 35 | 150 | 95 | 160 | Yes |
| 28 | 18 | 105 | 65 | 95 | No |

**Feature**:

* Individual attributes or columns in the dataset used for making predictions.
* **Example**: Age, BMI, Blood Pressure, Glucose Level, Insulin Level.

**Label**:

* The target variable or the output you want to predict.
* **Example**: Has Diabetes (Yes/No).

**Prediction**:

* The output generated by the model based on input features.
* **Example**: The model predicts "Yes" or "No" for "Has Diabetes".

**Outlier**:

* A data point that differs significantly from other observations.
* **Example**: If a patient has a Glucose Level of 200 in this dataset, it might be considered an outlier.

**Test Data**:

* A subset of the dataset used to evaluate the performance of the model.
* **Example**: A few rows from the dataset (e.g., 2 rows) separated for testing.

**Training Data**:

* A subset of the dataset used to train the model.
* **Example**: The remaining rows of the dataset after separating the test data.

**Model**:

* An algorithm or a mathematical representation trained on the dataset to make predictions.
* **Example**: A decision tree classifier predicting diabetes.

**Validation Data**:

* A subset of the dataset used to tune the hyperparameters of the model.
* **Example**: Another separate subset of data used during training.

**Hyperparameter**:

* Parameters that are set before the training process begins and control the training process.
* **Example**: The learning rate, the depth of a decision tree.

**Epoch**:

* One complete pass through the entire training dataset.
* **Example**: If the model trains on the entire dataset once, it completes one epoch.

**Loss Function**:

* A function that measures how well the model's predictions match the actual labels.
* **Example**: Mean Squared Error (MSE) for regression, Cross-Entropy Loss for classification.

**Learning Rate**:

* A hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated.
* **Example**: A learning rate of 0.01.

**Overfitting**:

* When a model performs well on training data but poorly on test data.
* **Example**: A model that memorizes the training data instead of learning general patterns.

**Underfitting**:

* When a model performs poorly on both training and test data.
* **Example**: A model that is too simple to capture the underlying patterns in the data.

**Regularization**:

* Techniques used to prevent overfitting by adding a penalty to the loss function.
* **Example**: L1 and L2 regularization.

**Cross-Validation**:

* A technique for evaluating the model by partitioning the data into multiple subsets and training/testing the model multiple times.
* **Example**: 5-fold cross-validation.

**Feature Engineering**:

* The process of creating new features or modifying existing features to improve model performance.
* **Example**: Creating a new feature like "BMI x Glucose Level".

**Dimensionality Reduction**:

* Techniques for reducing the number of features while retaining important information.
* **Example**: Principal Component Analysis (PCA).

**Bias**:

* The error introduced by approximating a real-world problem by a simplified model.
* **Example**: A linear model might have high bias if the true relationship is non-linear.

**Variance**:

* The error introduced by the model's sensitivity to small fluctuations in the training data.
* **Example**: A model with high variance might perform well on training data but poorly on test data due to overfitting.