| | Generative AI Consortium(Ltd)  **AI/ML Internship: Assignment 1**  **Name: HEMAARTHI G** | | --- | | **Email:** hemaaarthi0212@gmail.com |   **Dataset** | | | | | |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Year**  **Built** | **Make** | **Model** | **Mileage (Km)** | **Engine Size(L)** | **Condition** | **Resale Price (INR)** |
| 2005 | Toyoto | Corolla | 64000 | 1.8 | Good | 550000 |
| 2010 | Honda | Civic | 12800 | 2.0 | Excellent | 800000 |
| 2000 | Ford | Fiesta | 25600 | 2.6 | Good | 550000 |
| 2015 | Hyundai | Elantra2 | 60000 | 1.6 | Fair | 950000 |
| 2020 | Chevrolet | Malibu | 80000 | 2.5 | Excellent | 500000 |

**Feature:** Individual independent variables that act like an input in your system.  
*Example:* Year, Make, Model, Mileage (km), Engine Size (L), Condition.

**Label:** Identification of raw data.  
*Example:* Resale Price.

**Prediction:** Project a probable dataset that relates back to original data.  
*Example:* For a new record in the dataset with Year=2019, Make=Hyundai, and Mileage=20000, the model might predict Resale Price=800000.

**Outlier:** Data that is unique/different from other data.  
*Example:* If there was a car with a resale price of 2000000 INR in this dataset, it would be considered an outlier.

**Test Data:** Ensure that the model works for the given testing data.  
*Example:* Records of id=4 and id=5.

**Training data:** Data that is used to train the model.  
**Example***:* Records from id=1 to id=3.

**Model:** Program that can make decisions from previously unseen datasets.  
**Example***:* Support Vector Machine (SVM), Neural Network.

**Validation Data:** Uses a sample of data that is with-held from training.  
**Example***:* Records of id=2 and id=3.

**Hyperparameter:** Parameters that are set before training a model and control the learning process.  
**Example***:* The learning rate in a neural network.

**Epoch:** Each time a dataset passes through an algorithm, it is said to have completed one epoch. Therefore, it refers to the one complete passing of training data through an algorithm.  
**Example***:* One passes through records of id=1 to id=3.

**Loss Function:** Quantifies the difference between predicted outputs of a machine learning algorithm and actual target values.  
**Example***:* Mean Square Error (MSE).

**Learning Rate:** Tuning parameter in an optimization algorithm that determines the step size at each iteration while moving towards a minimum of a loss function.  
**Example***:* Starting with a learning rate of 0.01 and reducing it by a factor of 0.1 every 5 epochs.

**Overfitting:** A behavior that occurs when the learning model gives accurate predictions for training data but not for new data.  
**Example***:* If the model perfectly predicts the resale prices of the training data but fails to predict prices in the test data accurately.

**Underfitting:** When a model is too simple and has not learned the patterns in the training data well and is unable to generalize well on the new data.  
**Example***:* If the model predicts all car prices around 500000 INR regardless of their features.

**Regularization:** Set of methods to reduce overfitting.  
**Example***:* Dropout in neural networks.

**Cross-validation:** Technique of resampling different portions of training data for validation on different iterations.  
**Example***:* K-fold cross-validation, where the dataset is divided into K subsets and the model is trained and validated K times.

**Feature Engineering:** Technique that leverages data to create new variables that aren’t in the training set.  
**Example***:* Converting the "Year" to the "Age" of the car.

**Dimensional Reduction:** Method of reducing variables in a training dataset used to develop machine learning models.  
**Example***:* Principal component analysis (PCA).

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

**Variance:** Changes in the model when using different portions of the training dataset.  
**Example***:* A complex model that changes significantly with small changes in the training data has high variance.