Name: Amudhavan M

Roll No: 22CSR014

Department: Computer Science of Engineering

# Assignment 1: Documentation on Machine Learning Terminologies

## Introduction

This document provides a comprehensive understanding of various key terminologies used in the field of Machine Learning. The explanations are supplemented with an example test dataset to illustrate each concept clearly.

## Example Test Data: Used Car’s Resale Price History

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Make Year | Brand | Variant | Mileage | Fuel | Transmission | Resale Price (INR) |
| 2015 | BMW | 520D | 80000 | Diesel | Automatic | 2500000 |
| 2016 | Audi | A6 | 92000 | Petrol | Automatic | 1900000 |
| 2018 | Mercedes Benz | E200 | 61000 | Petrol | Automatic | 3400000 |
| 2014 | Skoda | Superb | 95000 | Petrol | Automatic | 600000 |
| 2020 | Benz | E200 | 35000 | Petrol | Automatic | 12000000 |

## Terminologies

### Feature

**Definition:** Features are individual measurable properties or characteristics of a phenomenon being observed.

**Example:** In the provided dataset, features include Make Year, Brand, Variant, Mileage, Fuel, and Transmission.

### Label

**Definition:** The label is the output variable that the model aims to predict.

**Example:** In the dataset, the Resale Price is the label.

### Prediction

**Definition:** Prediction is the process of using the trained model to estimate the label for new data.

**Example:** Estimating the resale price of a car given its features.

### Outlier

**Definition:** An outlier is a data point that deviates significantly from other observations.

**Example:** The resale price of 12000000 for a Benz in 2020 is an outlier compared to other prices.

### Test Data

**Definition:** Test data is a subset of the dataset used to evaluate the performance of the model.

**Example:** A separate set of used car records not included in the training dataset.

### Training Data

**Definition:** Training data is the subset of the dataset used to train the model.

**Example:** The provided dataset can be used as training data.

### Model

**Definition:** A model in machine learning is a mathematical representation of a real-world process.

**Example:** A regression model predicting resale prices based on car features.

### Validation Data

**Definition:** Validation data is a subset of the dataset used to tune model hyperparameters.

**Example:** A portion of the dataset reserved to evaluate model performance during training.

### Hyperparameter

**Definition:** Hyperparameters are parameters that are set before the learning process begins.

**Example:** Learning rate, number of epochs.

### Epoch

**Definition:** An epoch is one complete pass through the entire training dataset.

**Example:** Training the model for 100 epochs.

### Loss Function

**Definition:** A loss function measures how well the model's predictions match the actual data.

**Example:** Mean Squared Error (MSE) for regression tasks.

### Learning Rate

**Definition:** The learning rate is 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

**Definition:** Overfitting occurs when a model learns the training data too well, including the noise, resulting in poor performance on new data.

**Example:** A model that performs well on the training data but poorly on test data.

### Underfitting

**Definition:** Underfitting occurs when a model is too simple to capture the underlying pattern in the data.

**Example:** A model that performs poorly on both training and test data.

### Regularization

**Definition:** Regularization techniques are used to prevent overfitting by adding a penalty for larger model coefficients.

**Example:** L1 and L2 regularization.

### Cross-Validation

**Definition:** Cross-validation is a technique for assessing how the model will generalize to an independent dataset.

**Example:** k-fold cross-validation.

### Feature Engineering

**Definition:** Feature engineering is the process of using domain knowledge to create new features from raw data.

**Example:** Creating a new feature “Car Age” from “Make Year”.

### Dimensionality Reduction

**Definition:** Dimensionality reduction techniques are used to reduce the number of features in a dataset.

**Example:** Principal Component Analysis (PCA).

### Bias

**Definition:** Bias refers to the error introduced by approximating a real-world problem, which may be too complex, by a simpler model.

**Example:** A model consistently predicting lower resale prices than actual.

### Variance

**Definition:** Variance refers to the error introduced by the model’s sensitivity to small fluctuations in the training dataset.

**Example:** A model's performance drastically changing with different training datasets.

## Conclusion

Understanding these key terminologies is crucial for anyone involved in machine learning. This document serves as a foundational guide to help you grasp these concepts through definitions and examples.