**Machine Learning terminologies**

|  |  |  |  |
| --- | --- | --- | --- |
| **Petal-Colour** | **Petal-Shape** | **No of Petal** | **Name of Flower** |
| Red | Shallow bowl | 5 | Rose |
| Pink | Cup shaped | 6 | Saccate |
| Blue | Tubular | 4 | Butterfly Pea |

**Feature:** An individual measurable property or characteristic of a phenomenon or data point. Used as input variables for machine learning algorithms. Crucial for effective pattern recognition, classification, and regression. Eg-Petal-Colour,

Petal-shape and No of Petals are features in the above data set.

**Label:** A label, also known as the target variable or dependent variable, is the output that the model is trained to predict. In supervised learning, labels are the known outcomes that the model learns to associate with the input features during training. Eg-Name of Flower.

**Prediction:** Refers to the output of an algorithm that has been trained on a historical dataset. The algorithm then generates probable values for unknown variables in each record of the new data. The purpose of prediction in machine learning is to project a probable data set that relates back to the original data. This helps organizations predict future customer behaviours and market changes. Essentially, prediction is used to fit a shape as closely to the data as possible. Eg- Using petal colour, petal shape, no of petals we can predict name of the flower.

**Outlier:** An outlier is essentially a statistical anomaly, a data point that significantly deviates from other observations in a dataset. Outliers can arise due to measurement errors, natural variation, or rare events, and they can have a disproportionate impact on statistical analyses and machine learning models if not appropriately handled.

**Test Data:** A test dataset is a collection of data points that the model hasn’t seen during its training process.For example, ifa model is to recognize different types of dogs. You will feed it a large collection of images with labelled dog breeds (training data). The model learns the patterns and relationships between features like fur colour, ear shape, and body size to identify different breeds.

**Training Data:** It is used to train the machine learning model. Training data is the power that supplies the model in machine learning, it is larger than testing data. Because more data helps to more effective predictive models. When a machine learning algorithm receives data from our records, it recognizes patterns and creates a decision-making model.

**Model:** Machine learning models are computer programs that are used to recognize patterns in data or make predictions. Machine learning models are created from machine learning algorithms, which undergo a training process using either labelled, unlabelled, or mixed data. Different machine learning algorithms are suited to different goals, such as classification or prediction modelling, so data scientists use different algorithms as the basis for different models. As data is introduced to a specific algorithm, it is modified to better manage a specific task and becomes a machine learning model.

**Validation Data:** The validation is used to fine-tune the hyperparameters of the model and is considered a part of the training of the model. The model only sees this data for evaluation but does not learn from this data, providing an objective unbiased evaluation of the model.

**Hyperparameter:** Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process. These hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.

**Epoch:** An epoch is a complete iteration through the entire training dataset in one cycle for training the machine learning model. During an epoch, every training sample in the dataset is processed by the model, and its weights and biases are updated in accordance with the computed loss or error.

**Loss Function:** In machine learning, a loss function is used to measure model performance by calculating the deviation of a model’s predictions from the correct, “ground truth” predictions.

**Learning Rate:** The learning rate is a **hyperparameter** in machine learning that controls the step size at which the weights of a neural network are updated during training. It specifies the amount by which the model’s parameters are altered in the direction opposite to the gradient of the loss function.

**Overfitting:** Overfitting is an undesirable machine learning behaviour that occurs when the machine learning model gives accurate predictions for training data but not for new data. When data scientists use machine learning models for making predictions, they first train the model on a known data set. Then, based on this information, the model tries to predict outcomes for new data sets. An overfit model can give inaccurate predictions and cannot perform well for all types of new data.

**Underfitting:** Underfitting is another type of error that occurs when the model cannot determine a meaningful relationship between the input and output data. You get underfit models if they have not trained for the appropriate length of time on a large number of data points. Underfit models experience high bias they give inaccurate results for both the training data and test set.

**Regularization:** Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it. Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique.

**Cross-Validation:** Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data. It involves dividing the available data into multiple folds or subsets, using one of these folds as a validation set, and training the model on the remaining folds. This process is repeated multiple times, each time using a different fold as the validation set. Finally, the results from each validation step are averaged to produce a more robust estimate of the model’s performance. Cross validation is an important step in the machine learning process and helps to ensure that the model selected for deployment is robust and generalizes well to new data.

**Feature Engineering:** Feature engineering is the process of **transforming raw data into features that are suitable for machine learning models**. In other words, it is the process of selecting, extracting, and transforming the most relevant features from the available data to build more accurate and efficient machine learning models. Feature engineering involves a set of techniques that enable us to create new features by combining or transforming the existing ones. These techniques help to highlight the most important patterns and relationships in the data, which in turn helps the machine learning model to learn from the data more effectively.

**Dimensionality Reduction:** Dimensionality reduction is a technique used to reduce the number of features in a dataset while retaining as much of the important information as possible. In other words, it is a process of transforming high-dimensional data into a lower-dimensional space that still preserves the essence of the original data.

**Bias:** Bias is simply defined as the inability of the model because of that there is some difference or error occurring between the model’s predicted value and the actual value. These differences between actual or expected values and the predicted values are known as error or bias error or error due to bias. Bias is a systematic error that occurs due to wrong assumptions in the machine learning process.

**Variance:** In machine learning variance is the amount by which the performance of a predictive model changes when it is trained on different subsets of the training data. More specifically, variance is the variability of the model that how much it is sensitive to another subset of the training dataset. i.e. how much it can adjust on the new subset of the training dataset.  In simple words, variance tells that how much a random variable is different from its expected value.