**BASIC TERMS IN MACHINE LEARNING**

|  |  |
| --- | --- |
| **Year of Experience** | **Salary** |
| 1 | 5000 |
| 1.2 | 5500 |
| 1.4 | 6000 |
| 1.6 | 6500 |
| 1.8 | 7000 |

**The above dataset shows the employee experience in years and their corresponding salary.**

**Feature:** It is the measurable and individual quantity which is given as input to the machine learning model, through which it finds the output. Features can be numerical, categorial, binary, text. Selection of features is essential for accuracy of the prediction of ML models. Eg- year of experience in a feature in above dataset.

**Label:** It is the output that the model trying to predict. In supervised learning, the model is given label along with features or data, so that model knows for which type of pattern which type of label is given, simply models learns to identify patterns for the given label. Eg- Here the salary is the outcome to be predicted so it is label.

**Prediction:** It is the method which specifies the outcome of machine learning models using historical data. Simply, it the process of predicting output using historical data. In many areas artificial intelligence is making a big difference in predicting future using historical data. Eg- The prediction done by model using above dataset like salary(7500) for 2 years of experience

**Outlier:** It is the data point where there is much difference from other data points in its values. These outliers can global ( has big variation ) or contextual ( not has big variation ). These outliers can significantly affect the accuracy of the model. Eg- If the above dataset contains the experience 7 yeas and its corresponding salary is 75000.

**Test Data:** It can be simply called as data used in testing the machine learning model. These are the datasets that are not used in training or completely different from data used in training. Through this test data, models accuracy can be found.

**Training Data:** It can be simply called as data used to train the model. These training data is generally larger when compared to test data. Training data is very crucial in machine learning models because it is through which model learns patterns.

**Model:** It can be simply called as set of programs which is used in pattern detection and determining the output. Machine learning is completely dealt with models. Models can be supervised, unsupervised, semi supervised and reinforcement models.

**Validation Data:** It is the data used to tune the model. It is the subset of training data. It is different from both training data and test data.

**Hyperparameters:** These are the parameters that are set explicitly by user, which is used in controlling the learning process of the model. Eg- The value of K in K Nearest Number.

**Epoch:** It can be simply called as training the model completely using the training dataset one time. Eg – If training dataset has 100 sets of data, then training the model with 100 sets of data once is called as one epoch.

**Loss Function:** It can be simply called as the function used to measure the deviation in model’s prediction with actual data or It is the mathematical function used to measure the match between model’s prediction and actual data. Eg- Mean squared error loss – loss function calculated the mean of the squared deviation or difference from actual data.

**Overfitting:** It can be simply called as model works well on trained data( seen data ) and does not perform well on test data( unseen data ). Overfitting models not have more accuracy.

**Underfitting:** It can be simply called as model does not works well on both seen data and unseen data. The model cannot understand the patterns accurately, which results in inaccurate prediction of the model.

**Regularization:** It is the technique used to reduce the chance of overfitting by reducing the magnitude of the parameters. This results in a good machine learning model.

**Cross Validation:** It is the technique which involves of separating the data sets into different subsets and while training the model, using one set as validation set and other remaining sets as training sets, next time while training the model, another set is used as validation set and other remaining sets used as training datasets.

**Feature Engineering:** It is the process of selecting appropriate data from datasets and transforming it as suitable to train the machine learning model. In other word it can be said as creating or transforming the data into a most suitable feature.

**Dimensionality Reduction:** It is the process of reducing the number of features and still maintaining the important data as possible or it can be simply said as reducing from higher dimensionality to lower dimensionality by preserving the important data.

**Bias:** It is the type of error occurred due to deviation in prediction and actual data. It is the difference between the predicted values and actual data.

**Variance:** It is the deviation in predicted values or outcomes between different datasets. These bias and variance are reducible errors in machine learning.