**ML TERMINOLOGIES**

***STUDENT SCHOLARSHIP DATA***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***STUDENT ROLLNO*** | ***HIGH SCHOOL TYPE*** | ***PERCENTAGE*** | ***FAMILY INCOME*** | ***SCHOLARSHIP AWARDED*** |
| 22ADR001 | Govt | 78 | 150000 | Yes |
| 22ADR002 | CBSE | 88 | 1500000 | No |
| 22ADR003 | Govt | 92 | 100000 | Yes |
| 22ADR004 | Matric | 90 | 800000 | No |
| 22ADR005 | Govt | 72 | 80000 | Yes |

1. **FEATURES**

* It is an individual measurable property.
* Features are also referred to as variables or attributes.
* Eg: Student rollno, high school type, percentage, family income.

1. **LABEL**

* The output you get from your model after training it is called a label.
* Eg: Scholarship awarded.

1. **PREDICTION**

* The output of the data after training is called as prediction.
* Eg: Predicting Yes or No for student’s scholarship data.

1. **OUTLIER**

* A data point significantly different from other data points in a dataset.
* They represent errors in measurement.
* Eg: If there was a student with percentage 100 and family income 10000.

1. **TEST DATA**

* It is another subset of original data, which is independent of the training dataset.
* Eg: Student rollno 22ADR004,22ADR005.

1. **TRAINING DATA**

* It is the subset of original dataset which is used to train the model.
* Eg: Student rollno 22ADR001,22ADR002,22ADR003.

1. **MODEL**

* Models are output by algorithms and are comprised of model data and a prediction algorithm.
* Model represent the program.
* Eg: Logistic regression is used to train on the training data.

1. **VALIDATION DATA**

* It is a subset of data used to tune model parameter.
* Eg: Using 15% of data to adjust the hyperparameter.

1. **HYPERPARAMETER**

* Parameter that are explicitly defined by user to control the learning process.
* Eg: Learning rate.

1. **EPOCH**

* One entire passing of training data through the algorithm.
* Eg: If the model trains for 5 epochs it passes through training data 5 times.

1. **LOSS FUNCTION**

* A function that measures the difference between the predicted labels and actual labels.
* Eg: Using cross-entropy loss to evaluate prediction.

1. **LEARNING RATE**

* It is a tuning parameter in an optimization algorithm that determines the step size during training process
* Smaller learning rate may be slower but more precise training.

1. **OVERFITTING**

* Occurs when the model tries to cover all the data points or more than the required data points present in the given dataset.
* The overfitted model has low bias and high variance.
* Eg: If the model perfectly predicts SCHOLARSHIP AWARDED for training data but fails on new students.

1. **UNDERFITTING**

* Occurs when the model is not able to capture the underlying trend of data.
* An underfitted model has high bias and low variance.
* Eg: Predicting No for all students regardless of their features.

1. **REGULARIZATION**

* It is a technique to prevent the model from overfitting by adding extra information.
* Eg: Adding L2 regularization to the logistic regression model.

1. **CROSS VALIDATION**

* Technique to evaluate the model by splitting the data into several folds and training or testing on different folds.
* Eg: Using 5-fold cross-validation to assess model performance.

1. **FEATURE ENGINEERING**

* Leverages data to create new variables that are not in the training dataset.
* Eg: Combine percentage and family income into a single ‘Academic and financial standing’ feature.

1. **DIMENSIONALITY REDUCTION**

* Reducing the features in a training dataset to develop ML models.
* Eg: Using PCA to combine Family income and High school type into fewer features.

1. **BIAS**

* Systematic error in the model’s predictions.
* It is the difference between actual and predicted values
* Eg: If the model consistently underestimates the probability of awarding scholarships.

1. **VARIANCE**

* The model’s sensitivity to small fluctuations in the training data.
* Eg: If the model’s prediction vary when trained on different subsets of data.