ASSIGNMENT – 1

STUDENT TEST DATA SET

Here is a sample test data of student information with the highest CGPA during the past years.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ROLL NUMBER** | **NAME** | **GENDER** | **STUDY YEAR** | **DEPARTMENT** | **CGPA** |
| 22CSR001 | ABINITHY | FEMALE | 2017 | CSE | 9.5 |
| 22CSR002 | AKSHAY | MALE | 2018 | CSE | 9.1 |
| 22ECR004 | BINDHU | FEMALE | 2020 | ECE | 9.2 |
| 22ITR005 | CHARLIE | MALE | 2021 | IT | 9.5 |
| 22CSR006 | DANIEL | MALE | 2022 | CSE | 9.4 |

**Feature**

* Individual measurable properties or characteristics used as inputs to the model.

Eg: 'Roll Number', 'Name', 'Gender', 'Study Year', and 'Department'.

**Label**

* The output variable that the model aims to predict. Eg:'CGPA'.

**Prediction**

* The output value generated by the model when given a set of input features.

Eg: predicting the CGPA of a student.

**Outlier**

* A data point that deviates significantly from the rest of the dataset. If a student in this dataset had a CGPA of 4.0, it would be considered an outlier.

**Test Data**

* The subset of the dataset used to assess the performance of the model. Eg: If we use 20% of our data for testing, one of these rows could be part of the test data.

**Training Data**

* The subset of the dataset used to train the model. Eg: Remaining 80% of the dataset used to teach the model.

**Model**

* The mathematical representation of the relationship between features and labels.

Eg: regression model predicting CGPA.

**Validation Data**

* A separate subset used to fine-tune the model parameters

**Hyperparameter**

* These are parameters that are not learned from the data but set before training the model. Eg: Learning rate, number of epochs, etc.

**Epoch**

* One complete pass through the entire training dataset. Eg: training on all student data once is one epoch.

**Loss Function**

* A method to measure how well the model's predictions match the actual labels.

**Learning Rate**

* A hyperparameter that controls how much the model's weights are updated during training.

**Overfitting**

* When a model performs well on training data but poorly on test data. Eg:if our model perfectly predicts the training data but fails on new data.

**Underfitting**

* When a model is too simple to capture the underlying pattern of the data. Eg: a model that always predicts the average CGPA.

**Regularization**

* Techniques to prevent overfitting by adding a penalty to the loss function.Eg: L1 and L2 regularization.

**Cross-Validation**

* A method to assess the model’s performance by partitioning the data into subsets, training the model on some subsets, and validating it on others.

**Feature Engineering**

* The process of creating new features or modifying existing ones to improve model performance. Eg: creating a 'Years Since Enrollment' feature from 'Study Year'.

**Dimensionality Reduction**

* Techniques to reduce the number of features in the dataset. Principal Component Analysis (PCA) is a common method.

**Bias**

* The error introduced by approximating a real-world problem, which might be complex, by a simpler model.

**Variance**

* The error introduced due to the model's sensitivity to small fluctuations in the training set.