**VARSHINI .S**

**22CSR231**

**ML intern Assignment - 1**

**Dataset Table:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Student I** | **Hours Studied** | **Class Participation** | **Homework Completed** | **Final Grade(%)** |
| **001** | **10** | **High** | **8** | **95** |
| **002** | **7** | **Medium** | **6** | **78** |
| **003** | **5** | **Low** | **4** | **60** |
| **004** | **12** | **High** | **9** | **92** |
| **005** | **3** | **Low** | **2** | **50** |
| **006** | **8** | **Medium** | **7** | **85** |

**Machine Learning Terminologies:**

**Feature:**

* There are Individual measurable properties(e.g., Homework Completed ,Class Participation)were used as inputs to the model.

**Label:**

* The output variable that the model aims to predict (e.g., Final Grade).

**Prediction:**

* The output of the model after it were trained and using the features the model can predict the label(e.g., here predicting the final grade and Homework Completion).

**Outlier:**

* A data point that significantly from the rest of the data(for e.g., if student had a final grade of 100% despite studying in1 hour ,it consider as outlier).

**Test Data:**

* A subset of the dataset used to evaluate the performance of a trained model. It is not used in training (e.g., Student Ids 005 and 006 could be used as test data).

**Training Data:**

* A subset of the dataset used to train the model(e.g., Student IDs 001 to 004 could be used as training data).

**Model:**

* A mathematical representation of the relationship between the features and the label(e.g., a linear regression model predicting Final Grade based on Hours Studied ,Class Participation).

**Validation Data:**

* A subset of the training data used to tune the hyperparameters of a model. It helps in validating the model’s performance during training.(e.g., using student ids 003 and 004 could be used as training data).

**Hyperparameter:**

* Parameters set before the learning process begins, like the learning process begins, like the learning rate (e.g.,the number of hidden layers in a neural network).

**Epoch:**

* One complete pass through the entire training dataset.(e.g., if you train the model for 10 epochs, it means the model has seen the entire training dataset 10 times).

**Loss Function:**

* Measuring hoe the model’s predictions match the actual labels(e.g., using Mean Squared Error to measure the difference between predicted and the actual final grades).

**Learning Rate:**

* Controls how much the model’s weights are adjusted with respect to the loss gradient (e.g., a learning rate of 0.01 means the model updates weights slowly).

**Overfitting:**

* Occurs when the model learns the training data too well, including noise and details that don’t matter(e.g, overfitted model performs well on training but poor in new data’s).

**Underfitting:**

* Occurs when the model is too simple to capture the underlying pattern in the data (e.g, a model that performs poorly on both the training and test data).

**Regularization:**

* Techniques that prevent overfitting by adding or to the loss function(e.g., adding a penalty to the loss function for large coefficients in a linear regression models).

**Cross-Validation:**

* A method to ensure the model performs well by splitting the data into multiple parts and training several times (e.g -fold cross-validation involves splitting the data into 5 parts, training the model on 4 parts, and testing on the remaining part, repeating this process 5 times).

**Features Engineering:**

* Creating new features from existing ones to improve model performance (e.g., combining Hours Studied and Class Participation to create a new feature).

**Dimensionality Reduction:**

* Techniques that reduce the number of features while retaining important information(e.g., Homework Completed to two principal components).

**Bias:**

* An error introduced by simplifying made by the model. A high-bias model might make overly simple predictions(e.g., assuming final grades only depend on hours studied and ignoring class participation and homework completed).

**Variance:**

* The model’s sensitivity to small fluctuations in the training data(e.g., a high variance model may predict final grades accurately for training data but poorly for new unseen test data).