**Assignment 1**

**Case Study on Students Mark**

| **Student** | **Hours studied** | **Previous Exam Mark** | **Final Exam Mark** |
| --- | --- | --- | --- |
| A | 2 | 50 | 55 |
| B | 5 | 60 | 65 |
| C | 1 | 45 | 50 |
| D | 3 | 70 | 75 |
| E | 4 | 80 | 85 |

**Terminologies:**

* **Feature**: In this dataset, "Hours Studied" and "Previous Exam Mark" are features.
* **Label**: "Final Exam Mark" is the label that we aim to predict.
* **Prediction**: If we use the model to predict the "Final Exam Mark" for a student, the predicted value is the prediction.
* **Outlier**: If there was a student with "Hours Studied" as 100 and "Final Exam Mark" as 30, it would be an outlier.
* **Test Data**: If we reserve student E's data to test the model, it will be the test data.
* **Training Data**: The data of students A, B, C, and D used to train the model.
* **Model**: A regression algorithm that uses "Hours Studied" and "Previous Exam Mark" to predict "Final Exam Mark"
* **Validation Data**: If we split student D's data to tune the model, it will be the validation data.
* **Hyperparameter**: Parameters like the learning rate or the number of epochs set before training the model.
* **Epoch**: If the model processes the entire dataset of students A, B, C, and D once, it completes one epoch.
* **Loss Function**: Mean Squared Error (MSE) might be used to measure the difference between predicted and actual "Final Exam Mark".
* **Learning Rate**: If the learning rate is 0.01, it determines the step size during each iteration of model training.
* **Overfitting**: If the model performs well on students A, B, C, and D but poorly on student E, it might be overfitting.
* **Underfitting**: If the model performs poorly on both training and test data, it is underfitting.
* **Regularization**: Techniques like L2 regularization can be used to prevent overfitting.
* **Cross-Validation**: Dividing the dataset into folds and using different folds for training and testing to evaluate model performance.
* **Feature Engineering**: Creating a new feature like "Improvement in Previous Exam Mark" by calculating the difference between the "Final Exam Mark" and "Previous Exam Mark"
* **Dimensionality Reduction**: Using techniques like PCA to reduce the number of features if there were more features.
* **Bias**: If the model consistently predicts low scores, it might have high bias.
* **Variance**: If the model's predictions vary significantly with different training datasets, it might have high variance.