Training Day 6 Report:

Date: *2 July,2025(Wednesday)*

Location:*PG Block HPC Lab*

Guided by:*Training Instructors (Classroom-Based)*

Main Objective:

To understand practical data preprocessing for machine learning, including handling missing values, scaling and normalization, and implementing train-test split using Python libraries.

Summary of the Day's Work

Today's session focused on cleaning a real-world dataset using Python (pandas, numpy, scikit-learn), handling missing data with mean/median/mode, scaling and normalizing numeric features, and preparing the dataset for machine learning with an 80-20 train-test split.

Topics/Areas Covered:

- Data Preprocessing in Machine Learning.
- Handling Missing Values (NaN)
- Data Scaling using MinMaxScaler
- Normalization
- TechniquesTrain-Test Split (80% train, 20% test)
- Python Libraries: Pandas, NumPy, Scikit-learn

Concepts Learned:

- Missing values were handled using mean, median, and mode.
- MinMaxScaler was used to scale features between 0 and 1.
- Normalization helped bring data into the same scale.
- Pandas makes it easy to clean and explore tabular data.

- Learned how train-test split is used to separate data for ML models.

Tools / Platforms Used

- Google Colab / Jupyter Notebook
- Python 3.x
- Libraries: pandas, numpy, sklearn

Tasks Performed:

- Created a CSV file (student_data.csv) with columns: Name, ID, Math, Science, English.
- Inserted 5–10 rows, with some missing (NaN) values.
- Loaded CSV into Pandas DataFrame.
- Counted missing values using df.isnull().sum().
- Filled missing values using mean, median, and mode.
- Displayed updated dataset after cleaning.
- Performed MinMax scaling on numeric columns.
- Applied normalization to bring data into 0–1 range.

Observations / Reflections

This was my first time working with real dataset cleaning. I liked how pandas made it easy to handle NaN values and scale features. It felt like real data science work.

Key Takeaways

- Learned important preprocessing steps for machine learning.
- Gained confidence working with real datasets and popular Python libraries.
- Understood how proper cleaning and scaling improves data for ML models.