ASSIGNMENT-1

MODULE-1(GROUP TASK)

BUILD A SIMPLE ML PROCESS FLOW

Define the Problem

Example:

- Predict house prices → **Regression**
- Detect spam emails → **Classification**
- Group customers → **Clustering**

✓ Clearly define:

- Input data
- Expected output
- Success criteria (accuracy, error rate, etc.).

Collect Data

- Gather relevant data from:
  - Databases
  - CSV files
  - APIs
  - Sensors
  - Surveys

✓ More quality data = better model performance.

Data Preprocessing

This step cleans and prepares data.

Includes:

- Handling missing values
- Removing duplicates
- Encoding categorical variables
- Feature scaling (normalization/standardization)
- Splitting into:
    - Training set (usually 70–80%)
    - Testing set (20–30%).

Train the Model

- Feed training data to algorithm
- Model learns patterns
- Adjusts internal parameters

Evaluate the Model

Test using unseen data.

Common metrics:

- Accuracy
- Precision / Recall
- F1-score
- Mean Squared Error (MSE)

If performance is poor:

- Tune parameters
- Try another algorithm
- Improve data

Deploy the Model

- Integrate into:

    - Web app
    - Mobile app
    - Business system
- Model starts making real predictions

8️⃣ Monitor & Improve

- Monitor performance
- Retrain with new data
- Fix data drift issues

ML is **continuous improvement**, not one-time work.

This project demonstrates key ML concepts including:

- Data preprocessing
- Feature extraction
- Model training
- Model evaluation
- Deployment basics

In this project, text data (email content) is processed using **Natural Language Processing (NLP)** techniques and converted into numerical features using methods like **TF-IDF** or **Bag of Words**. A classification algorithm such as **Naive Bayes** or **Logistic Regression** is then trained on labeled data to learn patterns commonly found in spam messages.