Group Task - Module 3

Machine Learning: Concepts, Algorithms, and Applications

Build a simple mL Process Flow.

1)Problem Definition:

First, clearly define the problem.

Example:

Build a system that classifies emails as:

Spam

Not spam

2)Objectives:

- Improve email filtering accuracy
- Reduce unwanted messages
- Automatically detect suspicious content

Understanding the problem helps in selecting the correct mL approach.

3)Data Collection:

machine learning requires data.

sources of data:

- Email datasets
- Organization email logs.
- User-labeled spam data

Types of data:

- Text content of email
- Subject line.
- Sender information
- Links and attachments.

This data can be:

- structured
- unstructured.

3)Date Preprocessing :

now date is usually messy. we must clean it.

steps:

- Remove duplicates.
- Handle missing values .
- convert text to lower case.
- remove Stop words.
- remove special characters.
- Tokenization.
- Example: "Congratatatfons!!! you won a Pize!!!" ->"congratulations you won a prize."

4)Feature Extraction

- machine learning models cannot understand raw text data.
- we must convert text into numerical features.

Common Techniques:

- Bag of words
- TF-IDF
- Word Embeddings.

Features for spam detection:

- Number of suspicious words.
- Number of links
- Capital letter usage
- Email length.

5)Dataset splitting:

split the dataset into:

- Training data (70-80%)
- Testing data (20-30%).

This data can be:

➤ structured

➤ unstructured.

Training data used to train the model.

Testing data used to evaluate performance.

6)Algorithm selection:

since spam detection is a classification problem, we can use:

- Naïve Bayes
- logistic regression.
- Decision Tree
- Support vector machine.

7)model Training:

In this stage:

- The algorithm learns patterns from training data.
- It identifies relationships between features and labels.
- Example: if email contains 'lottery' + 'free' likely spam

8)model Testing:

After training:

- Test the model using unseen test data.
- Compare predicted labels with actual labels.

Evaluation metrics:

- Accuracy.
- Precision
- Recall
- F1-score
- confusion matrix

9)model Evaluation and Improvement:

- accuracy is low
- Improve feature extraction.
- Tune hyperparameters
- Try a different algorithm.
- collect more data

10)Deployment :

- Deploy model into real system.
- Integrate into email server.
- Automatically classify incoming emails

11)monitoring & maintenance:

- New Spam pattern appears
- model performance may drop.