**A purple circle with dots and lines

Description automatically generated**

**NAME:**

**Basit Abbas**

**ROLL NUMBER:**

**SU92-BSDSM-F23- 013**

**PROGRAM:**

**BS DATA SCIENCE**

**SUMBITTED TO:**

**SIR RASIKH**

***TASK : 06***

**Report on PAI\_Lab\_Tasks\_09 Notebook**

**1. Overview**

This notebook explores **text classification** using **Logistic Regression**. It includes key steps such as **preprocessing, exploratory data analysis (EDA), data splitting, model training, and prediction**.

**2. Key Sections**

**2.1 Preprocessing**

* The dataset (email\_text.csv) is loaded using **pandas**.
* Text data undergoes **tokenization, stopword removal, and lemmatization** using **NLTK**.
* Regular expressions (re.sub) are used to clean text by removing non-alphanumeric characters.

**2.2 Exploratory Data Analysis (EDA)**

* Although not explicitly detailed, EDA likely includes **basic statistical analysis** and **visualizations** to understand the dataset.

**2.3 Splitting Data**

* The dataset is split into **training and testing sets** to evaluate model performance.

**2.4 Model Training (Logistic Regression)**

* Logistic Regression is used for classification.
* Features and labels are processed before model training.

**2.5 Making Predictions**

* The trained model is tested on unseen data.
* Performance metrics (such as accuracy or confusion matrix) might be evaluated.