LAB 12

Naïve Bayes

You are tasked with building a (partial) Naïve Bayes classifier to distinguish between spam and not spam (ham) emails. The dataset provided requires preprocessing to handle various data issues.

Dataset:

The dataset `emails_dataset.csv` has the following columns:

Email Text: The text content of the emails.

Label: The label indicating whether the email is spam (1) or not spam (0).

Preprocessing Tasks:

1. Empty Cells: Remove rows with empty cells in either the "Email Text" or "Label" columns.
2. Data in Wrong Format: Convert the "Label" column to numeric format if needed.
3. Wrong Data: Check for any anomalies or wrong entries in the dataset and correct them.
4. Duplicates: Remove duplicate rows if any.

Naive Bayes Classification:

Implement a Naive Bayes classifier without using built-in libraries for Naive Bayes. Follow these steps:

1. Find probability of spam and not spam.
2. Find the unique words(vocabulary) in all of the emails.
3. Find probability of each word in vocabulary.
4. Compute conditional probabilities of each word. e.g. P (word = money| spam) and P (word = money| ham).