ASSIGNMENT 1

Assignment Topic Area: Naïve Bayes and Gaussian Naïve Bayes Classifier

Total Points: 50

Assignment Title: Spam Email Detection using Naïve Bayes and Gaussian Naïve Bayes

Problem Statement: In this assignment, you will implement a spam email classifier using two variants of Naïve Bayes models: Naïve Bayes (Multinomial Naïve Bayes) and Gaussian Naïve Bayes. The goal is to use a dataset of emails labeled as "ham" (non-spam) or "spam" to train and evaluate both models. You will apply standard machine learning techniques for text preprocessing, model training, evaluation, and comparison.

---

1. **Dataset Overview:**
   a) **Dataset Name:** Spam/Not-Spam Email Collection Dataset (Email Subject Only), Available on Kaggle ('spam_ham_dataset').
   b) **Features:**
      - **Text of the email subject:** Containing spam/non-spam relevant words/features.
      - **Label:** A binary label indicating whether the message is spam or ham.
   c) **Target Variable (y):**
      - **0 =** Ham (non-spam)
      - **1 =** Spam (spam)
   d) **Size:** 5171 email subjects (with 1499 spam subjects)

---

2. Assignment Instructions:

❖ **Data Preprocessing:**

Step 1.1: Load the Dataset

- Load the dataset into a Pandas DataFrame.
- Inspect the first few rows to understand its structure.

Step 1.2: Handle Missing Values

- Check for missing values in the dataset.
- Handle missing values by either dropping the rows with missing data or filling them appropriately.

Step 1.3: Split the Dataset

- **Split the dataset into features (X) and the target variable (y).**
    - **X:** Contains the message texts.
    - **y:** Contains the labels (spam or ham).

Step 1.4: Text Preprocessing

- **Tokenization:** Split the text into individual words or tokens.
- **Stop-word Removal:** Remove common but uninformative words such as "the," "and," "is," etc.
- **Lowercasing:** Convert all text to lowercase.
- **Optional:** You may use CountVectorizer or TfidfVectorizer from sklearn to convert the text into numerical features.

---

3. Modeling:

Step 2.1: Train-Test Split

- Split the dataset into training and testing sets using an 80-20 split (80% for training, 20% for testing).

Step 2.2: Multinomial Naïve Bayes (Naïve Bayes) Model

- Use sklearn.Naïve_bayes.MultinomialNB() to train a Naïve Bayes model.
- Train the model on the training set and make predictions on the testing set.

Step 2.3: Gaussian Naïve Bayes Model

- Use sklearn.Naïve_bayes.GaussianNB() to train a Gaussian Naïve Bayes model.
- Since text data is typically discrete, this model may perform poorly, but it serves as a baseline for comparison.

---

4. Evaluation:

Step 3.1: Evaluation Metrics Evaluate both models using the following metrics:

- **Accuracy:** The proportion of correct predictions. This is the overall success rate of the model.
- **Confusion Matrix:** Used to evaluate the performance of the classification models, with true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).
    - **True Positives (TP):** Spam correctly identified as spam.
    - **False Positives (FP):** Ham incorrectly classified as spam.
    - **True Negatives (TN):** Ham correctly classified as ham.

    - **False Negatives (FN):** Spam incorrectly classified as ham.
- **Precision, Recall, and F1-Score:**
    - Precision measures how many of the predicted spam emails were actually spam.
    - Recall measures how many actual spam emails were correctly identified.
    - F1-Score is the harmonic mean of precision and recall, providing a balanced metric.

---

5. Evaluation Criteria:

- **Correctness:** Whether the preprocessing, modeling, and evaluation steps are correctly implemented.
- **Clarity**: The organization and readability of the code.
- **Explanation:** Quality of the written discussion on model performance, including insights on the comparison between the models.
- **Technical Competence:** Proper use of machine learning and natural language processing (NLP) techniques.

---

6. Tools/Resources:

- **Scikit-learn:** Naïve Bayes, Gaussian Naïve Bayes
- **Text Preprocessing Techniques**: https://www.nltk.org/
- **Visualization with Matplotlib and Seaborn.**

7. Submission Guidelines:

a). Jupyter Notebook (25 Points):

- **File Format:** .ipynb
- **Content:**
    - Include all steps: data preprocessing, model training, evaluation, and visualizations.
    - Code should be well-commented and organized with markdown explanations.
    - Include results: accuracy, confusion matrix, classification results, ROC curve.

b). PDF Report (25 Points):

- **File Format:** .pdf
- **Length:** 2 pages (2-Colums) (IEEE format: Download IEEE Conf Template from IEEE Website)
- **Project Title, Authors Names**

- **Sections:**
    - **Abstract**: Briefly describe the proposed system for spam email detection.
    - **Introduction:** Briefly describe the spam detection problem, the Machine Learning models to be used, challenges of spam email detection, your contributions.
    - **Methodology:** Explain the dataset, preprocessing, and models (Multinomial Naïve Bayes and Gaussian Naïve Bayes).
    - **Results:** Present evaluation metrics (accuracy, precision, recall, ROC curve) and compare models.
    - **Conclusion:** Summarize findings and discuss model strengths and limitations.

c). IEEE Formatting:

- Use IEEE two-column format (Times New Roman, 10pt).
- **Keywords:** 3-5 relevant terms (e.g., "Spam Detection", "Naïve Bayes").

d). Submission Checklist:

- **Jupyter Notebook**: .ipynb with code and outputs.
- **PDF Report**: 2-page IEEE formatted report.

8. Self-Declaration: Please provide the self-declaration at the end of PDF report (Including the required two pages (two-columns), you can add an extra page for self-declaration). For example, you can follow the following:

"We, the undersigned, confirm that this assignment is a group project. Each member contributed to the tasks and deliverables as outlined below."

Contributions:

- **Member1:** Data preprocessing and feature extraction.
- **Member2:** Implemented Multinomial Naïve Bayes and evaluation.
- **Member3:** Implemented Gaussian Naïve Bayes and model comparison.
- **Member4:** Report writing and final compilation. (Adjust contributions for a group of 3)

Digital Signatures:

- Mem.1
- Mem.2
- Mem.3
- Mem.4 (no need for a group of 3)

N.B. [*The submission deadline will only be extended for exceptional circumstances, such as medical issues, with proper documentation or a recommendation from the Dean of Student Office.*]