Due Date: 08.12.2024 23:55

CENG 463 INTRODUCTION TO MACHINE LEARNING HOMEWORK 2

- Support Vector Machine
- Naïve Bayes
- Random Forest

In this assignment, you will work with a text-based dataset, the ChatDoctor Dataset (available here). Your primary goal is to process and balance this dataset for a multiclass classification task. Specifically, you will use the **reason** field in each example to classify the data into one of four categories: **disease**, **symptom**, **tests**, and **medication**. After preparing the dataset, you will implement and evaluate three machine learning models (SVM, Naive Bayes, and Random Forest) for this classification task.

Task 1: Data Preparation and Balancing

1. **Download and Load the Dataset**
    - **Task**: Download the dataset from the provided link and load it into a Pandas DataFrame.
    - **Subtasks**:
        - Display the first few rows to understand the structure of the dataset.
        - Check for missing values and clean the dataset as needed.
2. **Extract and Balance Classes**
    - The aim of this assignment is to classify each **reason** text into one of four categories: **disease**, **symptom**, **tests**, or **medication**.
    - **Task**: Create a balanced dataset where each class (disease, symptom, tests, and medication) has a sufficient and roughly equal number of examples. Ensure that the distribution of classes is as balanced as possible to improve model performance and prevent class imbalance issues.

Task 2: Model Implementation and Evaluation

For each model, train and evaluate it using the balanced dataset created in Task 1. Split your dataset into training and testing sets to ensure valid performance evaluation. Use **precision**, **recall**, and **F1-score** to assess the performance of each model.

1. **Support Vector Machine (SVM)**
    - **Data Preparation**: Scale the data using StandardScaler before training the model, as SVM requires scaled features for optimal performance.
    - **Model Training**: Train an SVM model on the training data.
    - **Evaluation**: Calculate and display precision, recall, and F1-score on the test set.
    - **Task**: Analyze the results and discuss why SVM may or may not be effective for this dataset.
2. **Naive Bayes**
    - **Model Training**: Train a Naive Bayes classifier using the same feature matrix and target variable.
    - **Evaluation**: Calculate precision, recall, and F1-score for Naive Bayes.
    - **Task**: Summarize the performance and discuss why Naive Bayes might be suitable or unsuitable for this text data.
3. **Random Forest**
    - **Model Training**: Train a Random Forest classifier using the training data.
    - **Evaluation**: Calculate precision, recall, and F1-score for Random Forest.
    - **Task**: Compare Random Forest's performance to SVM and Naive Bayes, and provide insights on why it performed as it did.

Task 3: Model Comparison and Justification

1. **Visualize Model Results**
    - For each model, create visualizations such as confusion matrices to compare true and predicted classes.
    - **Task**: Generate visualizations for each model to assess the alignment between actual and predicted classes.
2. **Model Performance Comparison**
    - Summarize each model's performance in a comparison table, listing precision, recall, and F1-score.
    - **Task**: Based on the comparison, discuss which model performed best and explain possible reasons. Consider factors like feature importance, model complexity, and interpretability.

You are expected to compare the performance of all models (with the performance metrics that you find in the task) as a table and answer that which model you think fits the data better and why.

Assignment Rules:

1. In this homework, no cheating is allowed. If any cheating is detected, the homework will be graded as 0, and no further discussion will be entertained.
2. You are expected to submit your homework in groups. Therefore, it will be sufficient if only one member of the group submits the homework.
3. You must upload a .txt file to MS Teams. In this file, include the link to your Google Colab notebook where you have done the project.
4. The .txt file must be named in the following format: group number, course code, and homework number. Example: **G01_CENG463_HW2**.
5. Please be aware that if you do not follow the assignment rules regarding export format and naming conventions, you will lose points.