Week 3: Model Training & Initial Evaluation

Task 1: Prepare Data for Training **(Tai)**
Goal: Format and split data for training/testing.
Steps:

- Load the .csv feature files from GitHub:
- data/american/mfcc_features/American_features.csv
- data/british/mfcc_features/British_features.csv
- data/vietnamese/mfcc_features/Vietnamese_features.csv
- Combine all .csv files into one dataset and add labels (0 = American, 1 = British, 2 = Vietnamese).
- Split the dataset into 80% training / 20% testing.
- Save as:
- data/training_data.csv
- data/testing_data.csv

Software: Google Colab, Pandas

Task 2: Train the Baseline Model **(James)**
Goal: Train an initial machine learning model for accent classification.
Steps:

- Load training_data.csv from GitHub.
- Choose a baseline model (Logistic Regression, SVM, or Random Forest).
- Train the model using extracted features.
- Save the trained model as:
- models/baseline_model.pkl
- Software: Google Colab, Scikit-learn

Task 3: Evaluate Model Performance **(Kyle & Wissam)**
Goal: Measure how well the model performs.
Steps:

- Load baseline_model.pkl and testing_data.csv.
- Evaluate using:
- Accuracy, Precision, Recall, and F1-score
- Confusion Matrix (to analyze misclassifications).
- Save results as:
- results/model_evaluation.txt

Software: Google Colab, Scikit-learn, Matplotlib