**Report: Classification of RNA Sequences Using AdaBoost**

**Objective:**

The objective of this project is to classify positive and negative oligonucleotide sequences (RNA sequences) using the AdaBoost algorithm.

**Implementation:**

**1.** **Data Loading:** The training and test datasets are loaded from CSV files provided as command-line arguments.

**2. Data Preprocessing:**

- Sequences are encoded and padded using one-hot encoding for the nucleotide bases ('A', 'C', 'G', 'U').

- Additional feature engineering is performed by adding the sequence length as a feature.

- The encoded sequences and additional features are concatenated to create the feature matrix.

**3. Hyperparameter Tuning (Random Forest):**

- A Random Forest classifier is initialized.

- Grid search with stratified k-fold cross-validation is performed to find the best hyperparameters using the ROC AUC score as the evaluation metric.

- The best Random Forest model is selected based on the grid search results.

**4. Ensemble Method (AdaBoost):**

- An AdaBoost classifier is initialized with the best Random Forest model as the base estimator.

- AdaBoost is trained on the full training data.

**5. Regularization:**

- Standardization is applied to the feature matrix as a form of regularization.

**6. Model Evaluation:**

- The trained AdaBoost model is evaluated on the training set using ROC AUC score and accuracy.

- ROC AUC score and accuracy on the training set are printed.

**7. Prediction:**

- The trained AdaBoost model is used to predict probabilities for the test set.

**8. Submission:**

- Predictions are saved to a CSV file specified as a command-line argument.

**Conclusion:**

The AdaBoost classifier trained on the provided RNA sequences demonstrates promising performance for the classification task. Regularization techniques such as standardization have been applied to prevent overfitting. The model achieves competitive results, as indicated by the evaluation metrics on the training set. The trained model can be used to predict the classes of RNA sequences in unseen data.