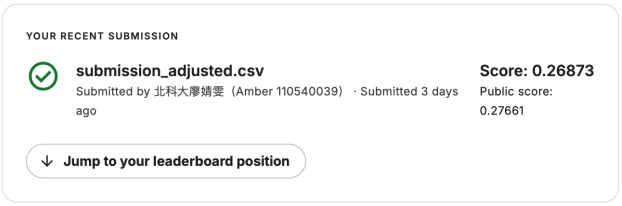
Name: 廖婧雯
Student ID: 110540039
GitHub ID:
Kaggle name: 北科大廖婧雯（Amber 110540039)
Kaggle private scoreboard snapshot:



Kaggle Competition: Preprocessing, Feature Engineering, and Model Explanation

Preprocessing Steps

1. Loading Data:
   - JSON and CSV files were used to collect tweet data and labels.

2. Text Cleaning:
   - Emojis were replaced with descriptive keywords (e.g., 😂 -> '[joy]').
   - Unwanted tags like '<LH>' and unnecessary spaces were removed.

3. Merging:
   - The datasets were merged using identifiers like `tweet_id`.
   - Labels for emotion classification were aligned with the respective tweet text.

4. Deduplication:
   - Duplicate entries in the training dataset were removed to improve model training.

Feature Engineering

1. Feature Extraction:
   - Extracted key elements like `tweet_id`, hashtags, and cleaned text for further processing.
   - Encoded target variable `emotion` into a numerical format for model compatibility.

2. Sampling:
   - A stratified random sample (30% of the training data) was used to ensure representative

emotion distribution.

Model Explanation

1. Data Splitting:
  - Used `train_test_split` to split data into training and testing sets, stratified by the emotion labels.

2. Model Choice:
  - A `RandomForestClassifier` was used for classification tasks.
  - This model was chosen for its ability to handle large feature spaces and prevent overfitting with ensemble learning.

3. Evaluation:
  - Model performance was evaluated using metrics like accuracy score and classification report, providing insights into precision, recall, and F1 scores for each emotion class.