Julius Baer Challenge:
Client Onboarding Process Simplified

Team **Daterminators**: Yihua Li, Jaiyi Chen, Varun Ghat Ravikumar, Tamanna Kumavat

Overview:

Our objective is to reduce the manual effort involved in onboarding clients by automating the verification of their KYC details through the following pipeline.

1. Data Extraction and Consolidation

- Extracted data from nested zipped folders.
- Created four separate CSV files, each corresponding to a different type of JSON file named: client_description_with_label.csv, account_form_with_label.csv, passport_data_with_label.csv, client_profile.csv.
- These CSVs consolidated various aspects of client information for easier downstream processing.

2. Data Exploration and Validation

- Conducted comparative analysis across the CSV files to detect:
  Discrepancies in client details like missing or inconsistent data points.
- Focused on the key fields for ex.Date of Birth, Passport Number and other Client Identifiers like FullName, address, phone number etc.

3. Rule-Based Filtering System

- Developed a rule-based pipeline to filter out clients who were rejected due to Incorrect or inconsistent data. Also checked mismatches in personal details across files (e.g., DOB in Client Description vs Passport Details) account_form_with_label.csv, passport_data_with_label.csv, client_profile.csv.
- These clients were excluded from modeling as their presence would not contribute meaningful labels/predictions.
- We were able to filter out 2875 clients because of the wrong/incorrect details after doing this step.

4. Ensured Data Consistency Between Client Profiles and Descriptions Using NLP

To reduce manual reading and verification effort during client onboarding, we automated consistency checks between the Client Profile and Client Description. Key fields: such as full name, age, marital status, job details, and savings amount—must match exactly, as discrepancies can lead to application rejection.

Technical Approach:

- Used the google/flan-t5-large model with prompt-based extraction to retrieve structured data from free-text client descriptions.
- Regularized outputs into dictionaries for comparison.
- Programmatically validated extracted fields against the Client Profile to flag mismatches.

5. Model Training with Multilayer Perceptron (MLP)

- Trained a 4-layer Multilayer Perceptron (MLP) model on the filtered and validated dataset.
- Input features were derived from filtered and NLP-extracted data, fields we used as features are: ['saving', 'inheritance', 'real_estate_value', 'total','investment_experience','preferred_markets','type_of_mandate','investment_risk_profile', 'label'].
- Labels were provided verified ground truth.
- On the trained dataset we were able to achieve 71% of accuracy on the filtered data.

6. Evaluation and Output

- We will evaluate the model on a provided designated evaluation dataset.
- Final predictions included:Client_IDs, Corresponding Predicted Labels.
- Results will be saved to an output file csv for the review.