**Advanced Data Cleaning with Fuzzy String Matching**

1. **Task Description**

This project focuses on advanced data cleaning techniques, specifically using fuzzy

string matching to perform data deduplication on the Titanic test dataset

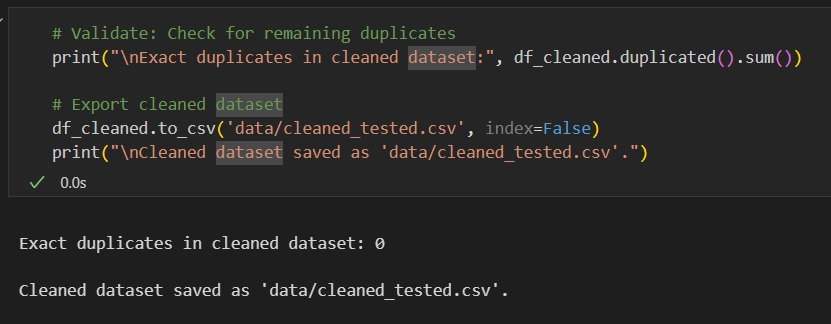
(tested.csv ) from Kaggle (https://www.kaggle.com/datasets/brendan45774/testfile).

The dataset contains 418 passenger records with columns such as PassengerId, Name, Age, Ticket, Pclass, and Embarked. The objectives are:

* Clean the dataset by handling missing values and standardizing text data.
* Identify near-duplicate entries in the Name column using fuzzystring matching.
* Remove duplicates (if any) and save the cleaned dataset as cleaned tested.csv.

The implementation is provided in a Jupyter Notebook (app.ipynb), which loads the dataset, processes it, and outputs a cleaned version after deduplication.

1. **Task Output Screenshot**



1. **Widget/Algorithm Used In Task**

* **Fuzzy String Matching (fuzzywuzzy)**: The fuzzywuzzy library with token\_sort\_ratio was used to identify near-duplicate names. It calculates Levenshtein distance, ignoring word order, to detect similarities (e.g., "Smith, Mr. John" vs. "Smith, Mr. Jon"). A 90% similarity threshold was applied.
* **Pandas**: Used for data manipulation, including loading the dataset, imputing missing values (Age with median, Embarked with mode, Cabin with "Unknown"), and standardizing the Name column.
* **Numpy**: Supported numerical operations, such as median calculation for Age.
* **Python-Levenshtein**: Enhanced fuzzywuzzy performance for string matching.
* **Jupyter Notebook**: The project was implemented in app.ipynb for interactive execution and result visualization.