**Internship Documentation – Netflix Dataset Preparation**

**1. Project Scope**

The objective of this milestone was to prepare the Netflix Kaggle dataset for analysis and potential recommendation system development. The dataset contains information about Netflix shows and movies such as title, director, cast, country, release year, rating, duration, genres, and description.

The primary scope included:

* Assessing and handling missing data.
* Removing duplicate or invalid records.
* Normalizing categorical fields (rating, country, and genres).
* Preparing a clean dataset that is consistent, reliable, and ready for downstream analytics or machine learning models.

**2. Success Metrics and KPIs**

To evaluate the success of the data preparation phase, the following **metrics** and **key performance indicators (KPIs)** were defined:

* **Data Completeness**:
  + % of missing values reduced to near zero for critical fields (e.g., title, rating, country).
  + Achieved: All records now have valid title, missing country values replaced with *Unknown*, and missing rating values replaced with *Not Rated*.
* **Data Consistency**:
  + Duplicate records removed based on unique identifiers (show\_id, title, director, cast, country, etc.).
  + Achieved: All exact duplicates were eliminated.
* **Data Quality**:
  + Cleaned dataset maintains original richness (≈ 8,800 rows).
  + Columns standardized for categorical encoding (ratings, countries, genres).
* **Preparedness for Analysis**:
  + Normalized features ready for model training (e.g., one-hot encoded vectors for rating, country, and genre list expansion).
  + Achieved: Genres tokenized into arrays and converted into feature vectors.

**3. Data Cleaning Achievements**

Several important cleaning steps were carried out on the dataset:

1. **Null Handling**
   * Rows with all-null fields were removed.
   * Rows with missing title were dropped, since title is a mandatory identifier.
   * Missing values in categorical fields were imputed with placeholder values:
     + country → *Unknown*
     + rating → *Not Rated*
2. **Duplicate Removal**
   * Duplicate rows were removed based on unique keys (show\_id, title, and other descriptive fields).
   * Ensures no show/movie is counted twice in analysis.
3. **Output**
   * Final dataset: **8,807 rows × 54 columns** (including original fields + encoded categorical features).
   * Dataset saved in both **Delta** and **CSV** formats for downstream analysis.

**4. Key Outcomes**

* Cleaned dataset is **complete, consistent, and structured**.
* Ready for exploratory data analysis (EDA), visualization, and modeling.
* Provides a strong foundation for building content-based or hybrid recommendation systems.