**Project Title:**

Predicting User Movie Preferences on Netflix

**Problem Statement:**

Netflix wants to improve its recommendation engine by predicting whether a user will like a movie or TV show based on their viewing history, demographics, and movie/series metadata (genre, cast, ratings, release year).

Your task is to perform the **end-to-end ML workflow** to build a predictive model.

**Steps:**

**1. Data Preparation**

* Handle missing values in user or movie data.
* Encode categorical variables such as genre, cast, and director.
* Normalize numerical features like movie duration, release year, and ratings.

**2. Data Validation**

* Split dataset into **training, validation, and test sets**.
* Ensure users in the test set also appear in the training set (to avoid cold-start leakage).
* Apply **stratified sampling** if necessary (e.g., like vs dislike balance).

**3. Feature Selection**

* Use **correlation analysis** to remove redundant features.
* Apply **feature importance methods** (Random Forest, XGBoost).
* Perform **dimensionality reduction** (PCA, SVD) for high-dimensional categorical data like cast/genre.

**4. Feature Engineering**

* Create **user-level features** such as:
  + Average rating given
  + Favorite genre
  + Most watched time of day
* Create **movie-level features** such as:
  + Popularity score
  + Average user rating
  + Trending flag
* Generate **interaction features** (e.g., user × genre preferences).

**5. Model Building**

* Train classification models such as **Logistic Regression, Random Forest, XGBoost, or Neural Networks**.
* Optionally explore **Collaborative Filtering** or **Matrix Factorization** approaches.
* Perform **hyperparameter tuning** with GridSearchCV/RandomizedSearchCV.

**6. Model Testing & Evaluation**

* Evaluate models using **accuracy, precision, recall, F1-score, ROC-AUC**.
* Compare baseline models with advanced recommendation models.
* Discuss **trade-offs** (e.g., precision vs recall in suggesting movies to users).

**Deliverables:**

* Preprocessed dataset with user and movie features.
* At least **3 trained models** with documented preprocessing pipelines.
* Comparison report of performance metrics.
* Final recommendation for improving Netflix’s recommendation system.

**Skills Tested:**

* Data preprocessing (handling missing data, encoding, normalization)
* Validation techniques (train-test split, cross-validation)
* Feature selection & engineering
* Model training (classification + recommendation models)
* Model evaluation and interpretation for recommendation systems