APR Assignment

Name: Sama Supratheek Reddy

Roll No: 2201CS62

Anime Score Classification Report

1. Objective
The goal of this project is to classify anime shows into score categories based on their features such as episodes, duration, genres, producers, studios, licensors, source material, and rating.
The target variable is Score_Class, bucketed as follows:

Score Range | Class
>= 8.0 | Very Good
7.0 - 7.9 | Good
6.0 - 6.9 | Average
< 6.0 | Low

2. Dataset Overview
- Source: anime.csv
- Initial dataset shape: 17562 rows x 35 columns
- Columns include: Episodes, Duration, Producers, Studios, Licensors, Genres, Source, Rating, Score, etc.

Target distribution after bucketing:
Average: 5300
Low: 3341
Good: 3232
Very Good: 548

Dropped score-related columns: ['Score', 'Score-10', 'Score-9', 'Score-8', 'Score-7', 'Score-6', 'Score-5', 'Score-4', 'Score-3', 'Score-2', 'Score-1']

3. Feature Preprocessing
Numeric features:
- Episodes -> numeric, missing values replaced with median.
- Duration -> converted to total minutes, missing replaced with median.

Categorical features:
- Source and Rating -> missing values replaced with "Unknown".

Multi-label features:
- Genres, Producers, Studios, Licensors -> split into lists.

- Top-K producers (40), studios (40), licensors (20).
- Genres with fewer than 30 occurrences removed (43 genres kept).

Final feature matrix shape: 12421 x 147
Target classes: ['Very Good', 'Good', 'Average', 'Low']

4. Train-Test Split
- Train size: 9936 samples
- Test size: 2485 samples
- Stratified split to maintain class distribution.

5. Model
- Classifier: SVM with RBF kernel
- Hyperparameters:
  - C = 10.0
  - gamma = scale
  - class_weight = balanced
- Pipeline includes preprocessing + SVM classifier.

6. Evaluation
Accuracy on test set: 0.6354

Confusion Matrix:

	Average	Good	Low	Very Good
Average	667	172	212	9
Good	165	408	35	39
Low	182	26	460	0
Very Good	4	60	2	44

Classification Report:
Average: Precision=0.66, Recall=0.63, F1-score=0.64, Support=1060
Good: Precision=0.61, Recall=0.63, F1-score=0.62, Support=647
Low: Precision=0.65, Recall=0.69, F1-score=0.67, Support=668
Very Good: Precision=0.48, Recall=0.40, F1-score=0.44, Support=110

Test class counts:
Average: 1060
Low: 668
Good: 647
Very Good: 110

Observations:
- Best performance for Average and Low classes.
- Very Good class has few samples, leading to lower recall (0.40).
- Overall accuracy is ~63.5%

7. Summary & Future Work

- Successfully preprocessed numeric, categorical, and multi-label features.
- SVM model captures patterns in episodes, duration, genres, and production details.

Future improvements:

- Hyperparameter tuning (GridSearchCV or RandomizedSearchCV)
- Try ensemble classifiers (Random Forest, Gradient Boosting)
- Reduce feature dimensionality (PCA or feature selection)
- Use embeddings for multi-label features to capture similarity