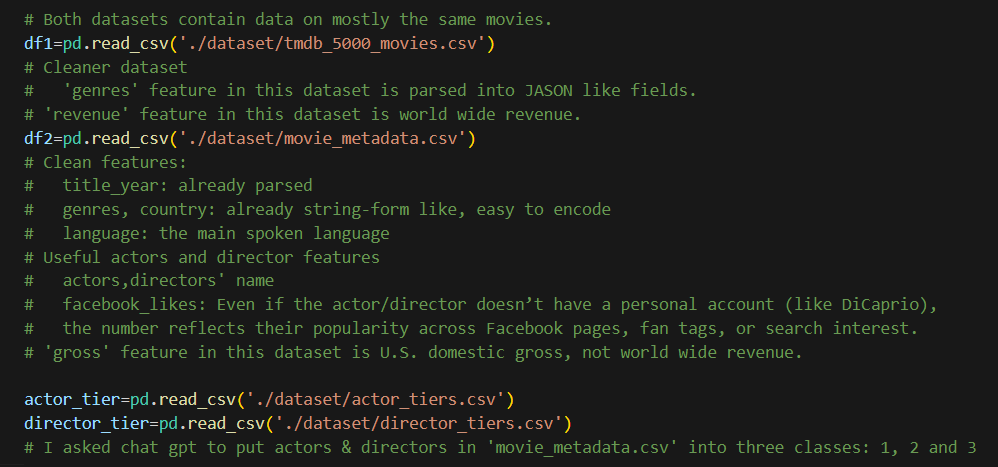
Data Mining AbdAlrahman Edrees

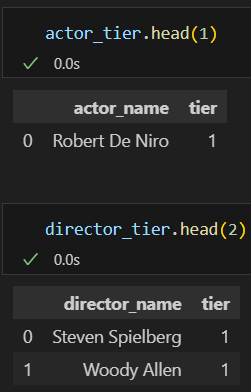
Last-decade American-Movies Data Mining

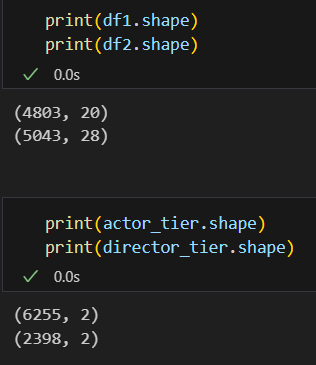
# Data Collection:



# Data Exploration:



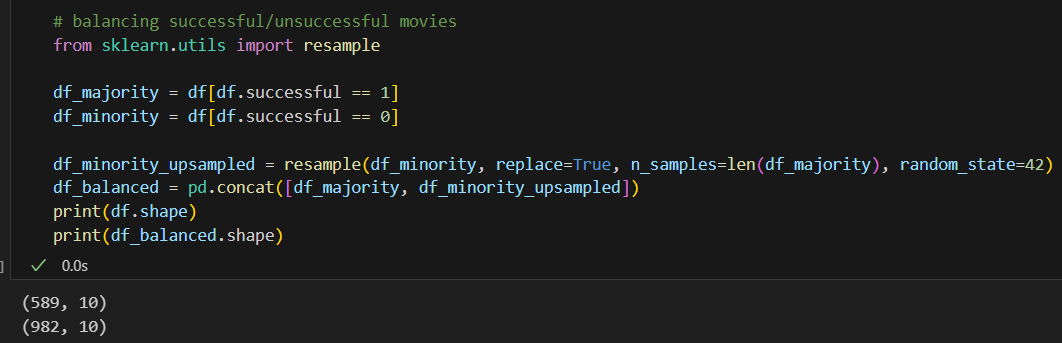




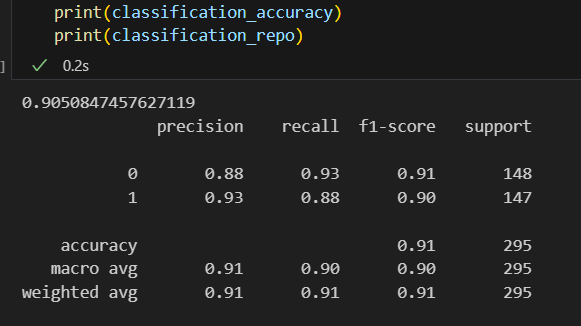
# Preprocessing:

1. We dropped unwanted features and merged the movies datasets (on= ‘title’).
2. We kept only English (language) American, produced after 2012 movies with budget > 100.000$ and runtime > 75 min -(movies from the same ecosystem).
3. We added ‘Succeed=1 if revenue>=budget’ feature.
4. Missing values treatment:  
   - budget, revenue: replaced with the mean  
   - content-rating: replace with the medium content-rating which is PG-13  
   - actor: replaced with a medium class actor  
   - genre: replaced with the most repeated genre
5. Outliers treatment:  
   - budget, revenue: Trimming.  
   - dropped rare genres and content-ratings.
6. Encoded genres, content-ratings, actor-names and director-names.

This is how the dataset looked like after that:

1. Balancing the dataset:

# Modeling:

Random Forest Classifier that predicts whether the movie will succeed or fail: