Data Collection and Preprocessing Phase

Date	23 September 2024
Team ID	LTVIP2024TMID24986
Project Title	Movie Box Office Gross Prediction using Machine Learning
Maximum Marks	2 Marks

Data Quality Report:

The Data Quality Report will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Quality Report:

Data Source	Data Quality Issue	Severity	Resolution Plan
Kaggle Dataset (Movie Data)	Missing values in the 'homepage', 'overview', 'release_date', 'runtime', and 'tagline' columns.	Moderate	Use appropriate imputation techniques (e.g., fill missing dates, mean/median imputation for numeric data).
Kaggle Dataset (Movie Data)	Categorical data in the 'genres', 'production_companies', 'production_countries', 'spoken_languages', 'cast', and 'crew' columns contain complex categorical data.	Moderate	Use one-hot encoding or label encoding for categorical columns. Consider handling multi-label columns (e.g., genres) using specialized techniques like multi-label binarization.

Kaggle Dataset (Movie Data)	Inconsistent formats in 'release_date' (e.g., different date formats)	Moderate	Standardize the date format and convert 'release_date' to datetime data type.
Kaggle Dataset (Movie Data)	Duplicate entries (possible duplicate movies based on 'title' and 'id')	Moderate	Check for duplicate records based on the 'title' and 'id' columns, and remove duplicates.