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 Collection Plan & Raw Data Sources Identification Report:

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan:

Section	Description
Project Overview	The machine learning project aims to predict box office revenue based on various features of movies, including budget, genres, cast, release dates, and more. Using a dataset with 23 columns such as budget, cast, revenue, and genres, the objective is to build a model that accurately forecasts a movie's box office gross, helping producers, distributors, and stakeholders in the film industry make informed financial decisions.
Data Collection Plan	● Search for datasets related to movie box office revenues, movie details (such as cast, crew, budget, and release dates), and production companies. ● Focus on datasets that include a wide variety of movies across different genres, release years, and production scales. ● Ensure that datasets have variables that are significant predictors of box office revenue. ● Collect additional datasets from sources like IMDb, TMDB, and Box Office Mojo.
Raw Data Sources Identified	The raw data sources for this project will include datasets obtained from IMDb, The Movie Database (TMDB), and Box Office Mojo. These platforms provide comprehensive details about movie productions, including key features like budget, revenue, cast, and genres. Additional datasets may also be collected from Kaggle or other open data repositories to enrich the prediction model.

Raw Data Sources Report:

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	The dataset comprises movie details such as budget, genres, cast, release date, and box office revenue. It also includes additional movie features like production companies and runtime, which may influence revenue predictions.	https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata	CSV	45.74 MB	Public