

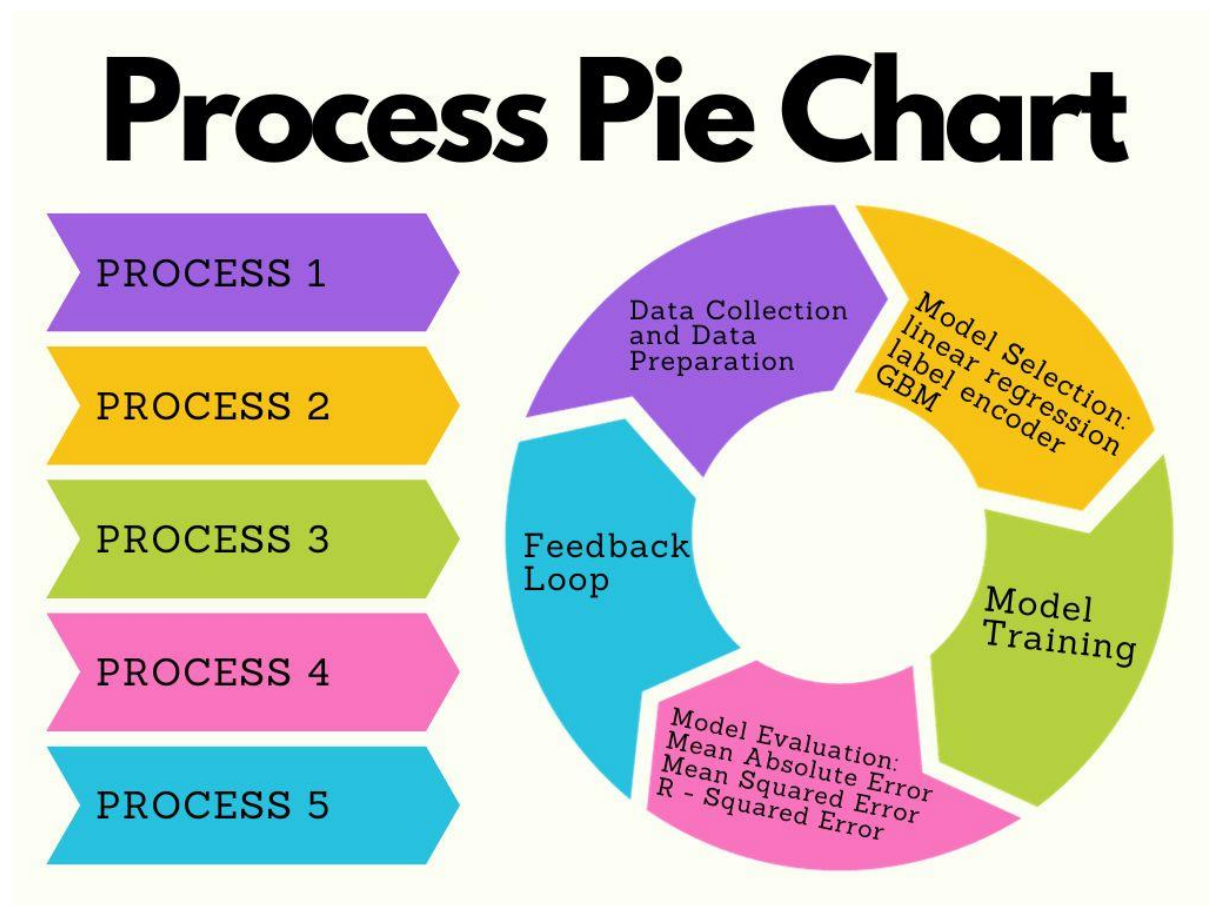
# IMDb SCORE PREDICTION

## (PHASE-2)

### Introduction:

Predicting IMDb scores for movies or TV shows is a complex task that often involves a combination of factors, including the quality of the content, marketing, critical reception, and audience reception. Estimate the popularity of movies. This estimation is based on certain features, including genre, premiere date, runtime, and language. In other words, to build a model that can predict how well a movie is likely to be rated on IMDb.

### Pie chart:



## Data Collection:

- **Movie/TV Show Information:** List all relevant details about the project, including title, release date, director, writer, cast, genre, budget, box office performance (if applicable), production company, and any notable awards or nominations.
- **Historical IMDb Data:** Mention any historical IMDb scores for similar movies or TV shows that will be used as benchmarks.

The dataset we used: <https://www.kaggle.com/datasets/luisortier/netflix-original-films-imdb-scores>

## Factors Affecting IMDb Scores

- **Cast Significance:**  
Discuss the importance of the cast, their star power, and acting talent.
- **Genre Impact:**  
Explain how the chosen genre may influence IMDb scores.
- **Budget and Production Values:**  
Describe the relationship between budget and production values and how they can impact audience perception.
- **Marketing and Promotion:**  
Discuss the role of effective marketing and promotion in generating audience interest.
- **Early Reviews and Previews:**  
Explain how critical and audience reactions from early reviews and preview screenings can provide insights.
- **Box Office Performance:**  
Mention whether box office performance is a relevant factor and how it correlates with IMDb scores.

## Data preparation:

Clean and preprocess the data, handling missing values, outliers, and encoding categorical variables

1. **Handling Missing Values:**
  - Decide on a strategy for dealing with missing data. Common options include:
  - Removing rows or columns with too many missing values.
  - Imputing missing values with the mean, median, mode, or a predictive model.
  - Using domain knowledge to infer missing values
2. **Data Cleaning:**
  - Address data inconsistencies and errors. This may involve:
  - Correcting typos and spelling mistakes.
  - Standardizing categorical variables.
  - Handling duplicates.
  - Removing irrelevant or redundant features.

## **Model selection:**

Once you have identified your features, you need to select a machine learning model to train. Some popular models for predicting IMDb scores include:

- Linear regression
- Label encoder
- Gradient boosting machines

## **Linear Regression:**

Linear regression is a widely used statistical method for modeling the relationship between a dependent variable (also called the target or outcome variable) and one or more independent variables (also called predictors or features). It is a fundamental technique in the field of machine learning and statistics, often employed for tasks like predicting numerical values, understanding correlations, and making forecasts.

## **Label Encoder:**

A label encoder is a preprocessing technique used in machine learning and data analysis to convert categorical or textual data into numerical values. Categorical data consists of labels or categories that do not have any inherent numerical meaning, such as colors, car brands, or job titles. Machine learning algorithms often require numerical inputs, so label encoding is used to represent categorical data in a format that can be used for model training.

## **Gradient boosting machines:**

Gradient Boosting Machines (GBM) is a powerful machine learning technique used for both regression and classification tasks. It's an ensemble learning method that builds a predictive model by combining the predictions of multiple individual models, typically decision trees. GBM has gained popularity in the field of machine learning due to its high predictive accuracy and robustness.

## **Model Training:**

Split our data into training and testing sets. Train our model on the training data. - Split our dataset into training and testing subsets (typically 70-80% for training and the rest for testing). This separation helps evaluate the model's generalization performance. - Train each selected model on the training data, using techniques like cross-validation to prevent overfitting.

## Model Evaluation :

Assess the performance of each model using appropriate regression evaluation metrics:

- **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual IMDb scores. Mean Absolute Error (MAE) is a commonly used metric in statistics and machine learning to evaluate the accuracy of a predictive model. It quantifies the average absolute difference between predicted values and actual values. In other words, it measures how close the model's predictions are to the true values in terms of magnitude, without considering the direction (overestimation or underestimation).

- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual scores, giving more weight to large errors. Mean Squared Error (MSE) is a widely used metric in statistics and machine learning to measure the average squared difference between predicted values and actual values. It quantifies the accuracy of a predictive model, with larger errors receiving more weight due to the squaring operation.

- **R-squared ( $R^2$ ):** Evaluates how well the model fits the data. A higher  $R^2$  indicates a better fit. R-squared, also known as the coefficient of determination, is a statistical measure used to evaluate the goodness of fit of a regression model. It provides information about how well the model explains the variation in the dependent variable. R-squared is a value between 0 and 1, with higher values indicating a better fit of the model to the data.

## Feedback Loop :

- Incorporate user feedback into the model's improvement cycle, allowing for adjustments to the feature set, model selection, and hyperparameters.

- Maintain a feedback loop to gather user ratings and incorporate them into the dataset for continuous model improvement.

## Conclusion:

Predicting IMDb scores for movies or TV shows is a multifaceted and challenging endeavor that requires a careful consideration of various factors and approaches. Predicting IMDb scores can be a valuable asset for filmmakers, production companies, and industry professionals seeking to gauge the potential success of their projects. However, it should be used in conjunction with other insights and feedback to make well-informed decisions in the ever-evolving world of entertainment.