**PHASE-1 Presentation**

**Project Name : Movie IMDb Score Prediction**

**Team Members:**

**1.Harish G**

**2.Hari Prasath S**

**3.Saktheeswaran C**

**4.Sangameshwar S**

**5.Rohithwarsan H**

**Movie IMDb Score Prediction**

**Problem Definition:**

The objective of this project is to develop a machine learning model that predicts IMDb scores for movies available on Films. This prediction will be based on various features such as genre, premiere date, runtime, and language. The goal is to create a model that accurately estimates the popularity of movies, helping users discover highly rated films that align with their preferences. The project encompasses several key steps, including data preprocessing, feature engineering, model selection, training, and evaluation.

**Design Thinking:**

**Data Source:**

To begin with, we need a dataset that contains information about movies, including the relevant features like genre, premiere date, runtime, language, and IMDb scores.The dataset will serve as the foundation for our machine learning model. It's essential to ensure that the data is comprehensive, accurate, and up-to-date.

**Data Preprocessing:**

The quality of the data is crucial for building a reliable predictive model. The following steps will be undertaken during data preprocessing:

**Data Cleaning:** Identify and handle missing values, outliers, and inconsistencies in the dataset.

**Data Transformation:** Convert categorical features (e.g., genre, language) into numerical representations using techniques like one-hot encoding or label encoding.

**Scaling:** Normalize or standardize numerical features to bring them to a similar scale, which can improve the performance of certain regression algorithms.

**Feature Engineering:**

Extracting relevant features from the dataset is vital for model accuracy.

This step involves:

Creating new features if needed, such as extracting the month from the premiere date or categorizing movies into sub-genres.

Selecting the most important features based on feature importance analysis.

**Model Selection:**

The choice of regression algorithm significantly impacts the model's predictive performance. We will consider various regression algorithms, including but not limited to:

Linear Regression: A simple model that assumes a linear relationship between features and IMDb scores.

Random Forest Regressor: A more complex ensemble model that can capture non-linear relationships and handle feature importance effectively.

The selection will be based on empirical testing and validation results.

**Model Training:**

Once the model is selected, it will be trained on the preprocessed dataset.

We will split the dataset into training and testing sets to evaluate the model's performance effectively.

The model will learn to predict IMDb scores based on the selected features.

**Evaluation:**

To assess the model's accuracy and performance, we will use several regression metrics, including:

Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual IMDb scores.

Mean Squared Error (MSE): Quantifies the average squared difference between predicted and actual IMDb scores.

R-squared (R2): Indicates the proportion of variance in IMDb scores that the model explains.

Cross-validation may also be employed to ensure the model's generalizability.

**Conclusion:**

This document outlines the approach for developing a machine learning model to predict IMDb scores for movies on Films. By following this structured process of data preprocessing, feature engineering, model selection, training, and evaluation, we aim to create a robust model that can help users discover highly rated films that align with their preferences. The choice of features and regression algorithms will be made systematically to optimize model performance. Regular evaluation and validation will be conducted to ensure the model's accuracy and effectiveness in predicting IMDb scores.