# IBM Naan Mudhalvan Project report

## Predicting IMDb Scores

DYNAMITE CODERS

Team Leader: Praveen Kumar S

Team Members: Mohan Kumar S , Rithik B , Thilak raj , Sri Hari V.

# IMDb Movie Score Prediction Project Documentation

## Table of Contents

5. **Phase 5: Project Documentation & Submission**

- Project Summary
- Data Documentation
- Model Documentation
- User Interface Documentation
- Challenges Faced
- Solutions Implemented
- Future Enhancements
- Scalability Considerations
- Final Review and Validation
- Project Submission

# Phase 1: Problem Definition and Design Thinking

## Problem Definition

The problem is to develop a machine learning model that predicts IMDb scores of movies based on features like genre, premiere date, runtime, and language. The objective is to create a model that accurately estimates the popularity of movies, helping users discover highly rated films that match their preferences. This project involves data preprocessing, feature engineering, model selection, training, and evaluation.

## Design Thinking Steps

### 1. Empathize:

a. Conducted user interviews and surveys to understand user preferences and pain points.
b. Explored existing movie recommendation systems to analyze strengths and weaknesses.

### 2. Define:

a. Defined the scope of the project, outlining key features and target audience.
b. Set specific goals, including accuracy benchmarks and user experience requirements.

### 3. Ideate:

a. Brainstormed potential features and data sources.
b. Explored different machine learning algorithms suitable for regression tasks.

### 4. Prototype:

a. Created wireframes and mockups for the user interface.
b. Developed a preliminary data preprocessing pipeline for feature engineering.

5. **Test:**
    a. Gathered feedback from potential users on the prototype.
    b. Iteratively refined the design based on user input.

## Phase 2: Innovation

## Introduction to Phase 2: Innovation

In Phase 2, the project transitioned from design thinking to innovation, focusing on transforming concepts into a functional solution. The phase included data preparation, model training, user interface development, and integration.

## Steps Taken in Phase 2

1. **Data Preparation and Integration:**
    a. Acquired movie data from various sources, cleaned it, and engineered features for machine learning.
    b. Harmonized data from diverse origins to create a solid foundation for the prediction model.
2. **Model Training:**
    a. Selected an appropriate machine learning algorithm based on its performance on the training and validation data.
    b. Fine-tuned the model's hyperparameters and trained it with a rich dataset.
3. **User Interface Development:**
    a. Crafted an intuitive and user-friendly interface allowing users to input movie features and receive IMDb score predictions.
4. **Model Integration:**
    a. Linked the trained model with the user interface, ensuring a cohesive and functioning system.
5. **Testing and Evaluation:**
    a. Conducted rigorous testing to assess the system's performance, usability, and user satisfaction.
    b. Gathered feedback for iterative development and refinement.

6. **Implementation and Scaling:**
    a. Prepared for full-scale deployment and devised strategies for scaling the system to accommodate growing user demands.
7. **Maintenance and Optimization:**
    a. Established plans for maintaining the system, addressing issues, and keeping the model and interface up to date.

## Phase 3: Development Part 1 - Loading and Preprocessing the Dataset

In this phase, the project started building the IMDb score prediction model by loading and preprocessing the dataset.

### Steps Taken in Phase 3

1. **Data Loading:**
    a. Retrieved movie dataset from IMDb, genre databases, premiere date records, and language information.
    b. Ensured data was available in suitable formats for analysis, such as CSV and JSON.
2. **Data Cleaning:**
    a. Removed duplicates and handled missing values using appropriate imputation techniques.
    b. Addressed outliers in the data that could skew predictions.
3. **Feature Engineering:**
    a. Selected relevant features such as genre, premiere date, runtime, and language for IMDb score prediction.
    b. Created new features, and encoded categorical data into numerical form using techniques like one-hot encoding.
4. **Data Splitting:**
    a. Divided the dataset into training, validation, and testing sets following the 70-15-15 rule.
    b. Randomly shuffled the data to ensure unbiased subsets.
5. **Data Visualization:**

   a. Utilized exploratory data analysis techniques to gain insights into the dataset.
   b. Calculated and displayed summary statistics for key features.

## Phase 4: Development Part 2 - Feature Engineering, Model Training, and Evaluation

In this phase, the IMDb score prediction model was further refined through advanced feature engineering, model training, and evaluation.

## Steps Taken in Phase 4

1. **Feature Engineering (Continued):**
   a. Applied feature transformations, created interaction terms, and explored dimensionality reduction techniques.
   b. Enhanced the dataset to capture more complex relationships.
2. **Model Training (Continued):**
   a. Confirmed the choice of the machine learning algorithm based on its performance on the training and validation data.
   b. Fine-tuned hyperparameters and utilized cross-validation techniques for model validation.
3. **Evaluation:**
   a. Evaluated the model's performance using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared ($R^2$) score.
   b. Ensured model interpretability and checked for signs of overfitting or underfitting.
4. **Model Fine-Tuning:**
   a. Made iterative adjustments to features, hyperparameters, and model complexity.
   b. Explored ensemble methods such as Random Forest and Gradient Boosting for improved accuracy.

## Phase 5: Project Documentation & Submission

## Steps Taken in Phase 5

1. **Project Summary:**
    a. Provided an overview of the project's objectives, key features, and technologies used.
2. **Data Documentation:**
    a. Documented data sources, data description, and preprocessing steps.
3. **Model Documentation:**
    a. Explained the model architecture, feature importance, and performance metrics.
4. **User Interface Documentation:**
    a. Described the UI design, layout, and user instructions for interacting with the system.
5. **Project Challenges and Solutions:**
    a. Outlined challenges faced during the project, including data quality issues and model limitations.
    b. Documented solutions implemented, including data cleaning techniques and algorithm optimizations.
6. **Future Enhancements:**
    a. Discussed potential improvements such as incorporating additional features and implementing advanced model architectures.
7. **Scalability Considerations:**
    a. Provided strategies for scalability, including optimizing database architecture and utilizing cloud resources.

## Conclusion:

This comprehensive document outlines the entire journey of the IMDb Movie Score Prediction project, from problem definition and design thinking to development, evaluation, and future considerations. It serves as a detailed record of the project's evolution and the methodologies employed, ensuring transparency and providing valuable insights for future projects and analyses.

## Declaration:

We have provided the colab notebook and all the required documents in Github repo. All our efforts are unique. This is to declare that we have completed the Nan Mudhalvan project successfully.

**GITHUB (Team lead) :**

**https://github.com/praveen0908/IBM_NAAN_MUDHALVAN_PROJECT-**

# THANK YOU