Project Report

Predicting IMDb Ratings Using Support Vector Regression (SVR) and Random Forest

Introduction

This project aims to predict IMDb ratings of movies using machine learning algorithms. We will employ Support Vector Regression (SVR) and Random Forest algorithms to build predictive models. The dataset includes features such as movie rank, year, duration, age limit, rating, number of ratings, Metascore, and description.

Dataset Overview

The dataset consists of the following columns:

- **rank**: The rank of the movie.
- **year**: The release year of the movie.
- **duration**: The duration of the movie in hours and minutes.
- **age_limit**: The age rating of the movie.
- **rating**: The IMDb rating of the movie.
- **numberof_ratings**: The number of ratings the movie has received.
- **Metascore**: The Metascore of the movie.
- **description**: A brief description of the movie.
- **name**: The name of the movie.

Data Preprocessing

Before building the models, we need to preprocess the data:

- Convert the **duration** column from the format "xh ym" to minutes.
- Extract numerical values from the **numberof_ratings** column, which is in the format "(2.9M)".
- Handle missing values, if any.
- Convert categorical columns (e.g., **age_limit**) into numerical values using encoding techniques.

Feature Engineering

Feature engineering involves creating new features or transforming existing ones to improve the model's performance. Possible steps include:

- Creating a feature that represents the movie's age by subtracting the release year from the current year.

Model Building

We will use two machine learning algorithms: Support Vector Regression (SVR) and Random Forest. Here are the steps involved in building these models:

Support Vector Regression (SVR)

1. **Normalization**: Normalize the features to ensure that all values are on a similar scale.
2. **Training**: Train the SVR model using the preprocessed data.
3. **Hyperparameter Tuning**: Use techniques such as GridSearchCV to find the best hyperparameters for the SVR model.

Random Forest

1. **Feature Importance**: Determine the importance of each feature to select the most relevant ones.
2. **Training**: Train the Random Forest model using the selected features.
3. **Hyperparameter Tuning**: Use GridSearchCV to optimize the hyperparameters.

Model Evaluation

Evaluate the performance of the models using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. Visualize the results using plots to compare the actual vs. predicted ratings.

Visualization

To visualize the results, we will create scatter plots comparing the actual IMDb ratings with the predicted ratings from both models. Using matplotlib library.

Conclusion

Summarize the findings and discuss the performance of the models. Highlight any challenges faced during the project and suggest potential improvements for future work.