# -Film Recommendation and IMDB Prediction System

**Group - 8**

* Vaishavi Bommagani
* Neelesh Bagal
* Akshaya Narsingu
* Sai Chanikya Eagam

**Motivation:**

The motivation behind this project stems from the evolving landscape of the entertainment industry, where digital platforms have become the primary medium for content consumption. As users are inundated with a vast array of film choices, there is a growing need for a personalized and intelligent film recommendation system. Traditional methods of selecting films often fall short in providing tailored suggestions, and this project seeks to address this gap by leveraging the power of machine learning. By doing so, it aims to enhance the user experience and streamline the process of discovering new and compelling films.

In the digital era, the sheer abundance of cinematic content, coupled with the rise of streaming platforms, has led to a paradox of choice for users. While the availability of diverse content is a boon, it has become increasingly challenging for viewers to navigate through this vast sea of options and discover films that align with their unique tastes. Traditional methods of recommendation often rely on broad categorizations, overlooking the intricate nuances of individual preferences. This project is motivated by a desire to alleviate the decision fatigue faced by users, offering them a curated selection of films tailored to their specific likes and dislikes.

Moreover, the advent of advanced technologies, such as machine learning, provides an unprecedented opportunity to transform the way we engage with films. The motivation is not just about making recommendations; it's about creating a cinematic journey that resonates on a personal level, fostering a deeper connection between users and the diverse world of storytelling.

**Significance:**

The significance of the Film Recommendation and IMDb Prediction System lies in its ability to revolutionize how users interact with cinematic content. Personalized recommendations not only make the viewing experience more enjoyable for users but also contribute to the success of the film industry by connecting audiences with content that aligns with their preferences. Additionally, the IMDb Prediction System adds value to filmmakers, producers, and distributors by offering insights into a film's potential success and critical acclaim before its release. This predictive capability can influence decision-making processes within the industry, leading to more informed choices in content creation and distribution.

The significance of this project extends beyond individual user experiences. In an industry where success is often measured by box office numbers and viewer ratings, the Film Recommendation and IMDb Prediction System can serve as a catalyst for positive change. By connecting users with films that genuinely resonate with them, the system has the potential to increase viewer satisfaction, leading to greater retention on streaming platforms and increased revenue for filmmakers.

Furthermore, the IMDb Prediction System introduces a novel approach to decision-making in the film industry. Filmmakers, producers, and distributors face significant financial investments and risks with each project. The ability to forecast a film's potential success and critical acclaim empowers industry professionals to make informed choices, ultimately contributing to the creation and distribution of high-quality content.

In a broader context, this project aligns with the evolving paradigm of personalized digital experiences. As users increasingly seek customization in various aspects of their lives, from online shopping to content consumption, the Film Recommendation and IMDb Prediction System stands at the forefront of catering to this demand in the cinematic domain.

Ultimately, the significance lies in the transformative potential of this project, not only enriching the entertainment experiences of individuals but also influencing the dynamics of the film industry itself. As we witness the convergence of technology and storytelling, this system emerges as a pioneering force that has the capacity to redefine how films are discovered, enjoyed, and produced in our rapidly evolving digital landscape.

**Objectives:**

*Developing a Reliable Model:* The primary objective is to create a robust machine learning model capable of accurately predicting film quality based on various attributes. This involves identifying key features that significantly contribute to film quality and building models that effectively leverage these features.

*User-Centric Film Recommendations:* The project aims to develop a film recommendation system that considers user preferences and viewing history. The objective is to provide personalized suggestions, enriching the user experience and introducing viewers to films they might have otherwise overlooked.

*IMDb Prediction System:* Integrate regression models to predict IMDb ratings, offering valuable insights into a movie's potential success and critical acclaim before its release. This objective supports decision-making processes for filmmakers, producers, and distributors.

*Ethical Considerations:* Address ethical considerations and challenges associated with potential biases in recommendations. The project aims to implement transparency in the recommendation system, providing users with insights into why a particular film is recommended or predicted to have a specific rating.

*Adaptability to Dynamic Trends:* Develop models that can adapt to the dynamic nature of user preferences and ever-changing trends in the film industry, ensuring the IMDb Prediction System remains relevant and accurate over time.

**Features:**

*User Preferences Integration:* The system will consider individual user preferences and viewing history to tailor film recommendations on a more personalized level.

*IMDb Prediction:* The incorporation of regression models to estimate IMDb ratings based on relevant features, providing valuable insights for industry professionals.

*Transparency:* The system will be designed to provide transparency into the recommendation and prediction process, allowing users to understand why certain films are suggested or predicted to have specific ratings.

*User Feedback Integration:* Strategies will be devised to seamlessly integrate user feedback and reviews into the model, consistently enhancing its accuracy and relevance.

*Flexibility for Cultural Differences:* The system will be flexible enough to accommodate cultural and regional differences in film preferences, ensuring that recommendations and predictions resonate with diverse audiences.

*Scalability:* The system will be scalable to handle a large user base and adaptable to challenges such as cold-start problems, where there is limited information available about a new film or a new user in the system.

**Related Work:**

1. Netflix Recommendation System: Netflix, a leading streaming platform, has pioneered personalized content recommendations. Their recommendation system utilizes collaborative filtering, content-based filtering, and deep learning models to analyze user behavior and preferences. Research papers and articles on Netflix's recommendation algorithms provide valuable insights into the challenges and solutions implemented by a major player in the streaming industry.
2. MovieLens Collaborative Filtering: MovieLens, a movie recommendation service, has been the subject of extensive research in collaborative filtering techniques. The MovieLens dataset and collaborative filtering algorithms have been widely explored in academic literature. Studies on the effectiveness of collaborative filtering methods, such as user-item and item-item collaborative filtering, offer a foundation for understanding recommendation system dynamics.
3. IMDb Rating Prediction Models: Various studies have explored IMDb rating prediction models using machine learning techniques. These models often consider features such as genre, cast, director, and user reviews to predict IMDb ratings. Research in this area delves into the factors that contribute most significantly to a film's rating and the challenges associated with predicting user-generated ratings accurately.
4. Hybrid Recommendation Systems: Hybrid recommendation systems, combining collaborative filtering and content-based approaches, have gained attention for their ability to provide more accurate and diverse recommendations. Research exploring the integration of different recommendation techniques can offer insights into how hybrid systems can improve the overall recommendation accuracy and user satisfaction.
5. Explainable AI in Recommendation Systems: With the increasing importance of transparency in AI systems, research on explainable AI (XAI) in recommendation systems has emerged. Studies focus on developing models that not only provide accurate recommendations but also offer understandable explanations for the recommendations made. This aligns with the ethical considerations mentioned in the project proposal.
6. Cold-Start Problem Solutions: Addressing the cold-start problem, where new items or users have limited data available, is a significant challenge in recommendation systems. Research explores solutions such as content-based recommendations for new items, hybrid models, and leveraging demographic information to mitigate the impact of cold-start problems.
7. Cultural and Regional Variations in Recommendations: Understanding cultural and regional differences in film preferences is crucial for a global recommendation system. Research in this area explores methods to adapt recommendation algorithms to diverse cultural contexts, ensuring that the system provides relevant and resonant suggestions for users worldwide.
8. User Feedback Integration in Recommender Systems: Studies on the integration of user feedback and reviews into recommendation systems offer valuable insights into improving system accuracy over time. Techniques such as sentiment analysis on user reviews and incorporating implicit feedback contribute to the iterative refinement of recommendation algorithms.

**Dataset:**

*IMDb Ratings Dataset*: The IMDb Ratings dataset contains valuable information about movies, offering insights into their titles, release year, genre, duration, country of production, language, director, cast, and various attributes related to their critical and user reception. For the purpose of this project, the primary attributes deemed essential are ratings, votes, title, description, country, and language.

Attributes:

* imdb\_title\_id: Unique identifier for each IMDb title.
* title: The title of the movie.
* year :The year of release.
* genre :The genre of the movie.
* duration: The duration of the movie in minutes.
* country: The country of production.
* language: The language of the movie.
* director: The director of the movie.
* actors: The main actors in the movie.
* Description: A brief description of the movie.
* avg\_vote : The average user vote for the movie.
* votes : The number of votes the movie received.

*Netflix Titles Dataset*: The Netflix Titles dataset provides information about shows and movies available on Netflix, offering details about their titles, type, director, cast, country of production, date added, release year, rating, duration, listed genres, and a brief description. For the purpose of this project, the primary attributes considered crucial are ratings, title, description, country, and language.

 Attributes:

* show\_id: Unique identifier for each Netflix title.
* type: The type of content (Movie or TV Show).
* title: The title of the content.
* director: The director of the content.
* country: The country of production.
* release\_year: The year of release.
* rating: The content rating.
* duration: The duration of the content.
* description: A brief description of the content.

Numerical Features:

1. IMDb Ratings Dataset: year, duration, avg\_vote, votes.
2. Netflix Titles Dataset: release\_year, rating, duration,.

Categorical Features:

1. IMDb Ratings Dataset: imdb\_title\_id, title, genre, country, language, director, actors, description.
2. Netflix Titles Dataset: show\_id, type, title, director, country, description.

We have collected the IMDb Ratings and Netflix Titles datasets from Kaggle.

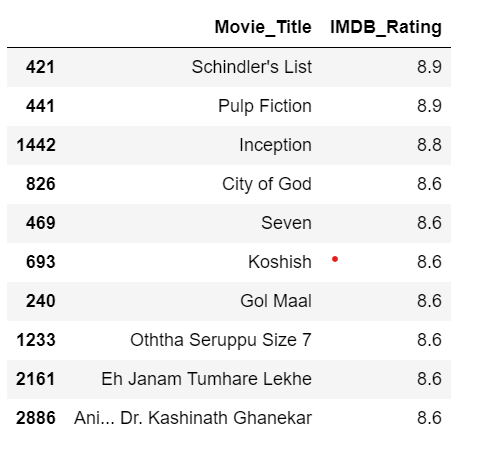
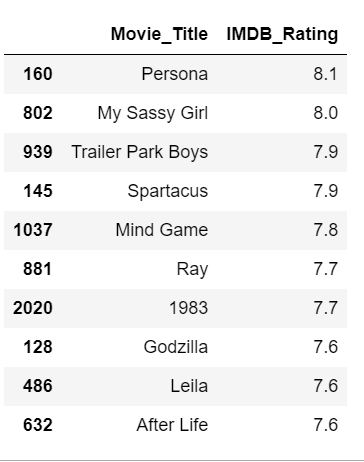
Netflix Titles dataset: <https://www.kaggle.com/shivamb/netflix-shows>

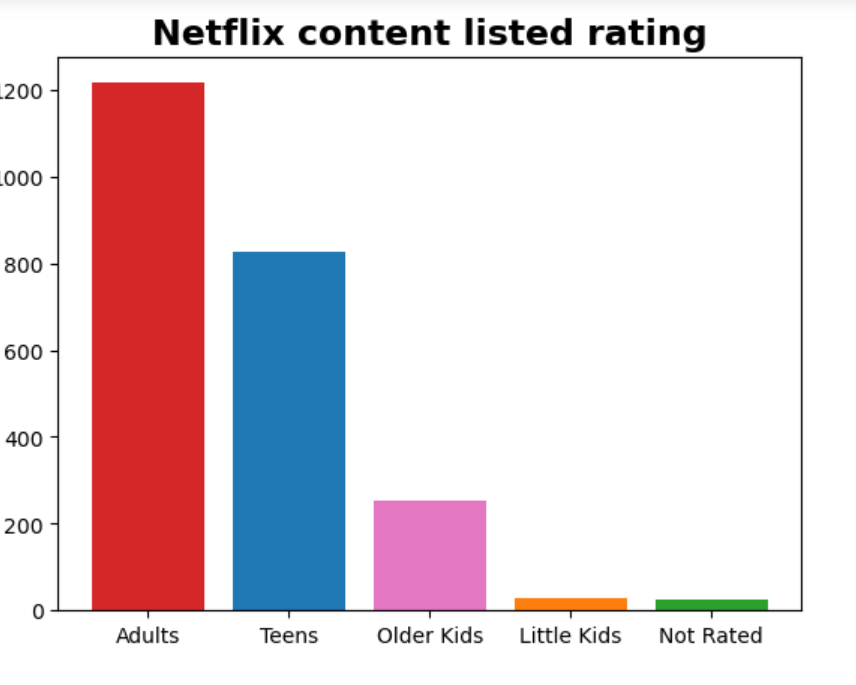
IMDb Ratings dataset: <https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset>

**2. Analysis (EDA):**

In the step basically we will visualize and analyse the data.

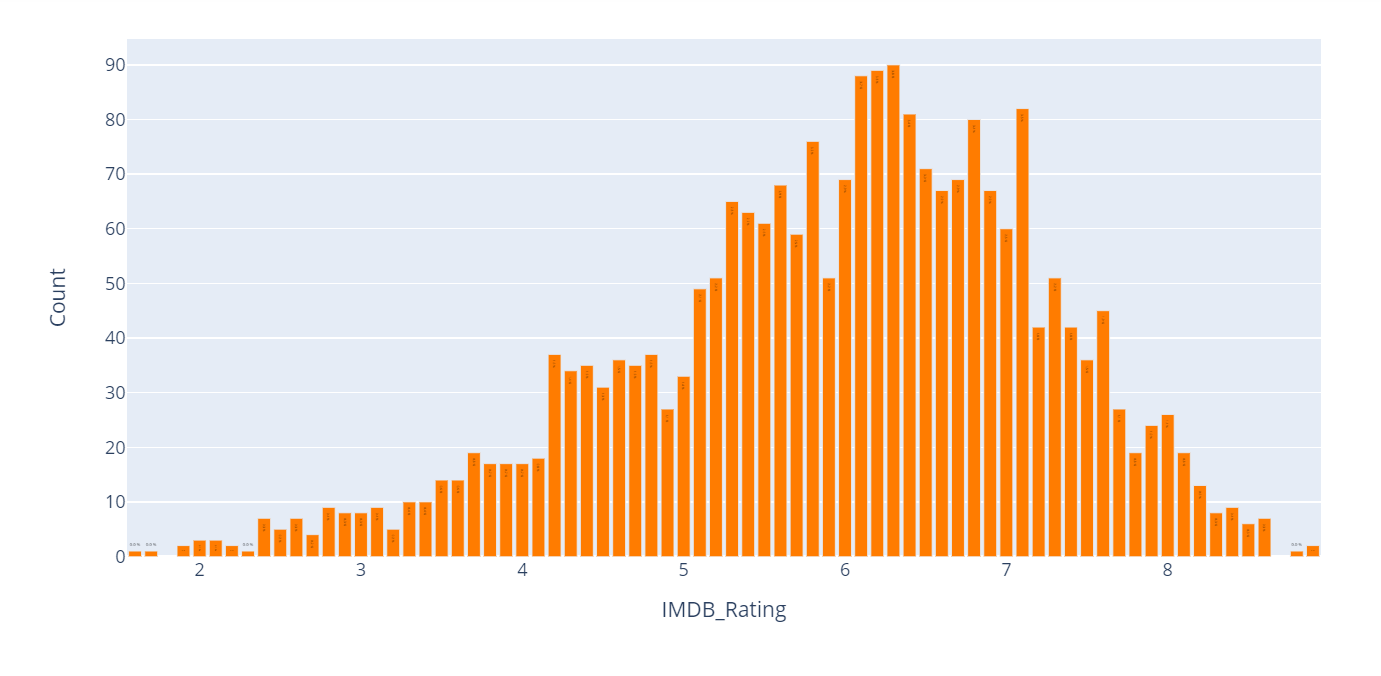
First we looked at the top 10 movies and the top 10 TV shows based on IMDb Rating

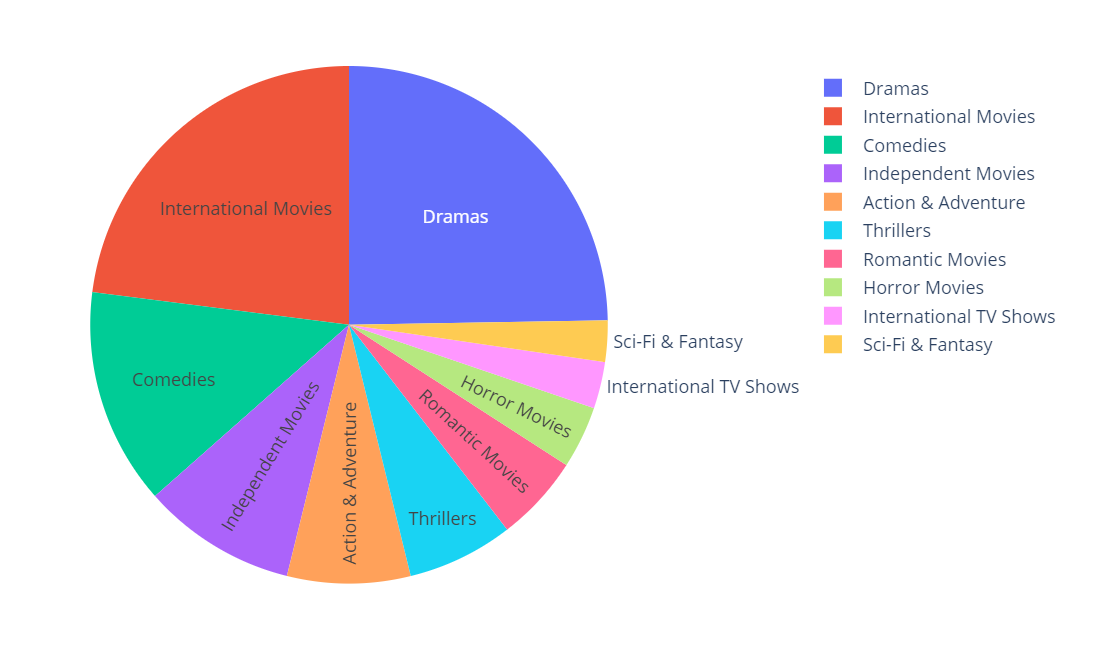
->We have plotted a bar graph to visualize to find which age group the Netflix content is more relevant. To visualize it we have plotted bar graph with rating value count against age groups.

As we can see, the content is mostly for adult audience (50%) which refers to older than 17 years (mature).

->Then we have plotted a 3D bar graph **from plotly for distribution of all content's ratings by their counts.**

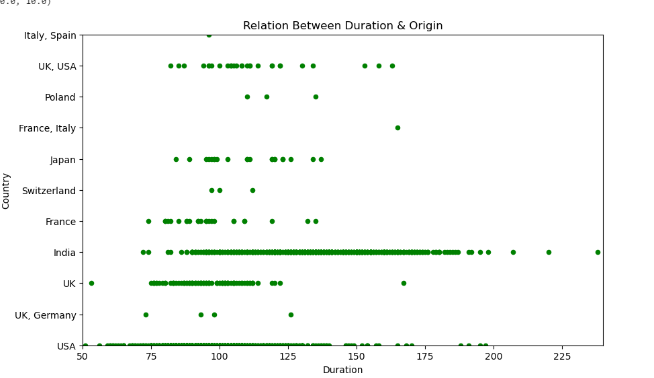


**->A pie-chart to visualize the distinct Genres distribution and their percentages.**

****

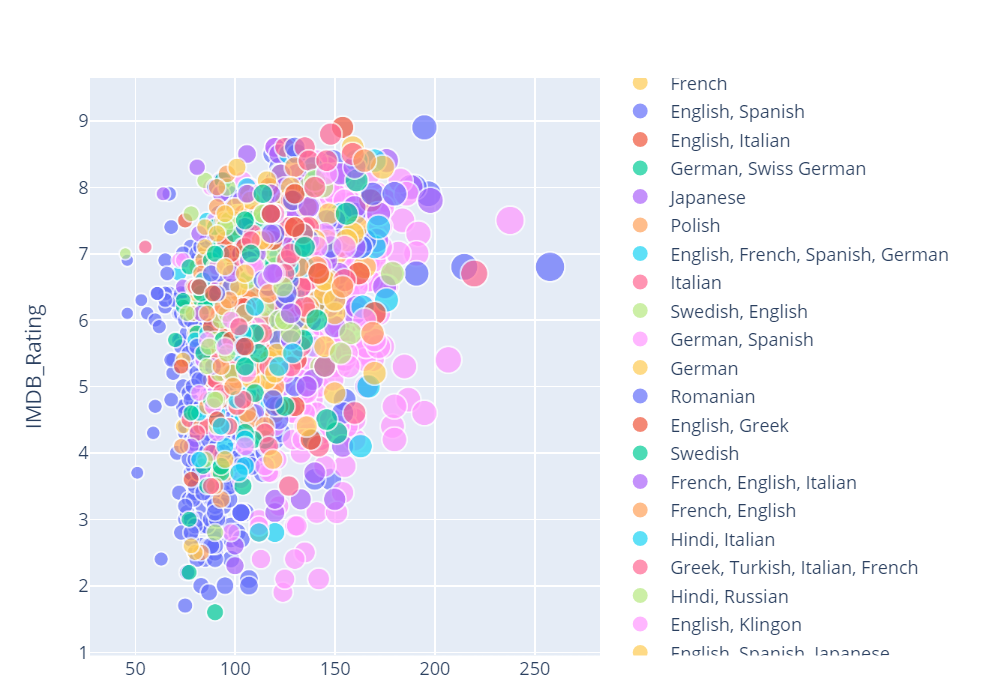
Dramas, International Movies, Comedies are the top 3 dominant genres with 25%,23%,13% percentages.

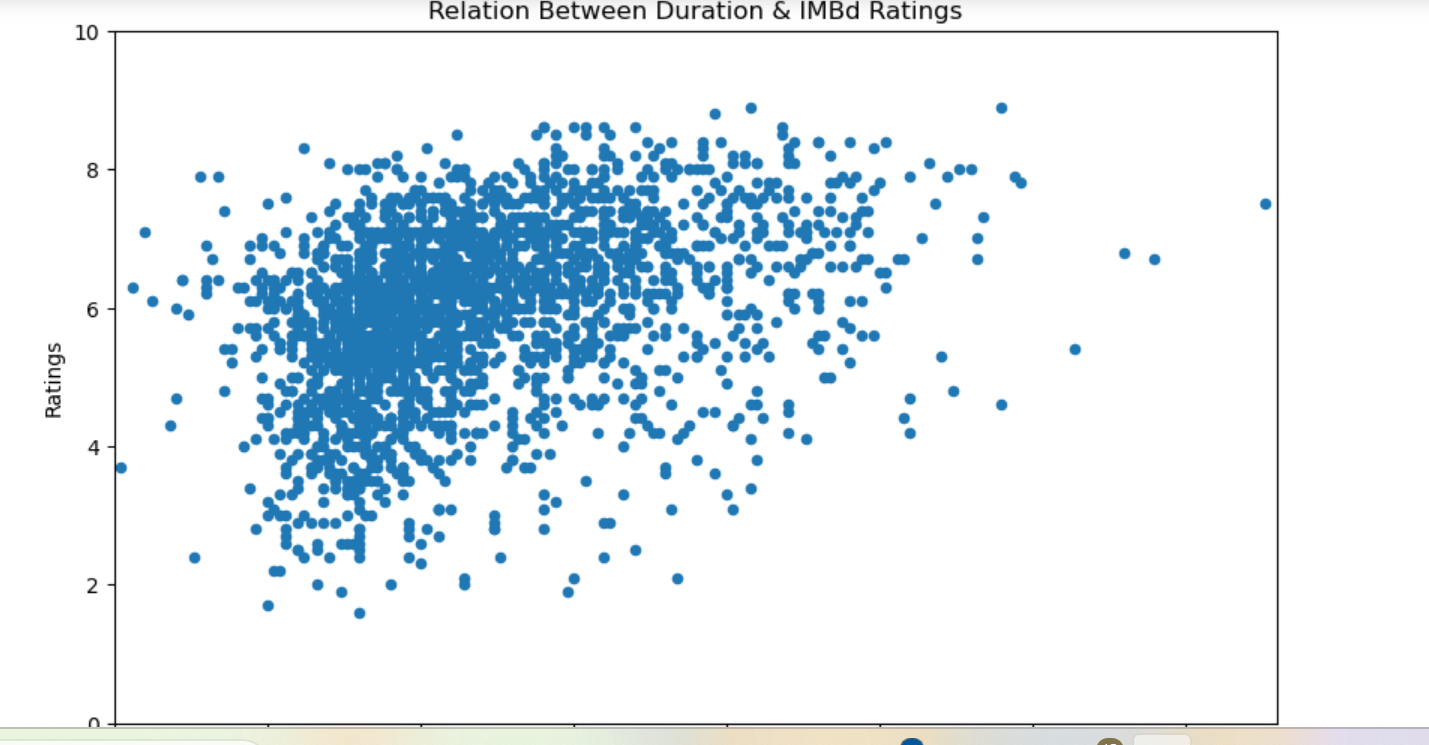
**->Scatter plots to show the relation between Duration and Country.**

****

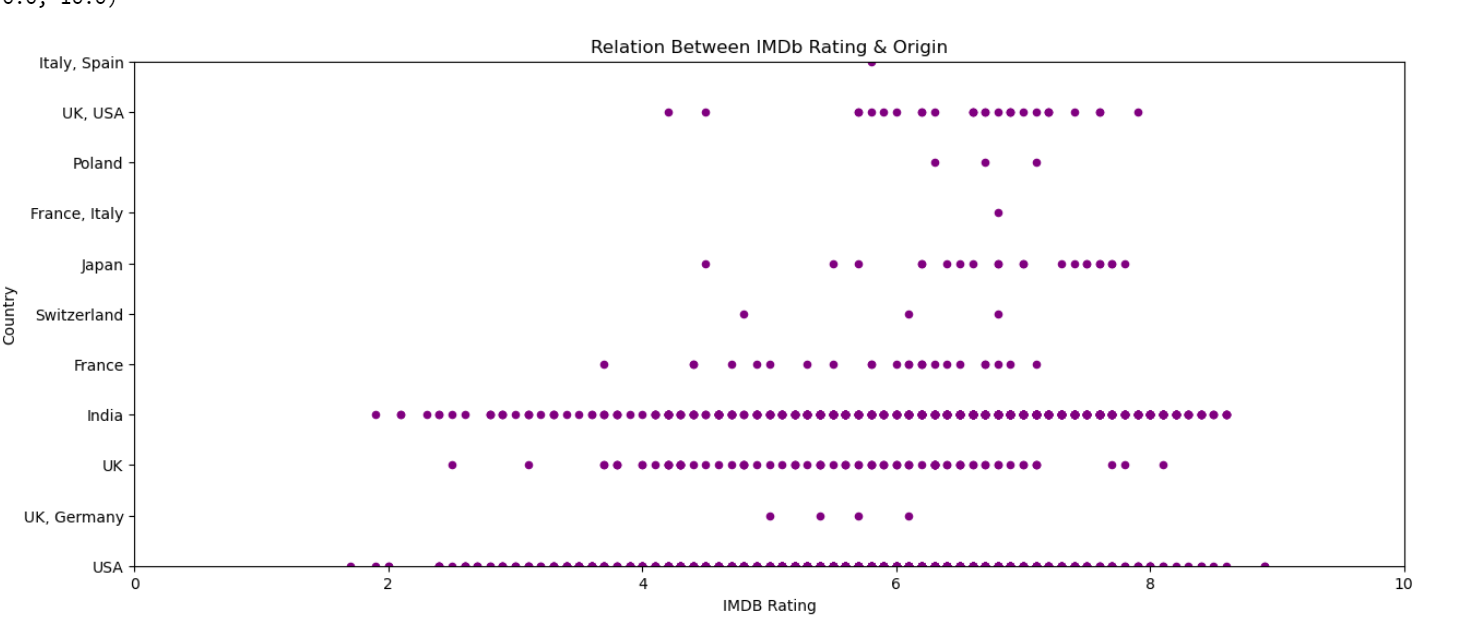
**Most movies with duration 180 and above are from India.**

**->Scatter plot between Duration and IMDb Rating.**

****

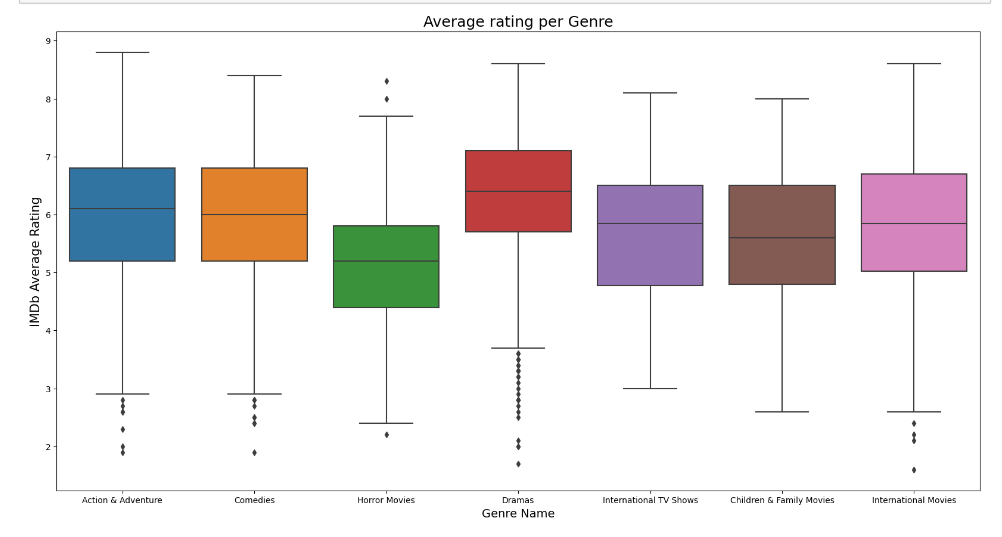
****

**->Scatter Plot between IMDb Rating and Country**

****

We can see that films with high IMDb Rating are from India and Usa.

->Box plot to analyse average rating across different generes.

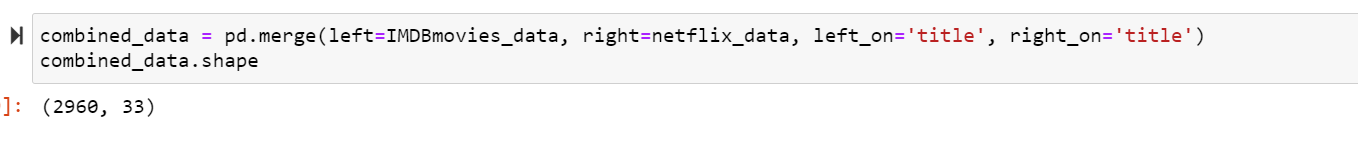


The rating distributions across genres show no notable differences; generally, they hover around 6. The sole exception is in Horror Movies, where a significant distinction is observed with an average rating of approximately 5.

**Detail Design of Methods:**

**Data Preprocessing:**

a.First Load the IMDb Rating dataset and the Netflix Titles dataset and merge both the datasets to one dataset based on the movie/TV show titles. Inorder to Recommend the Movie or TV Show we need to merge them.



Drop the Duplicate Values:

In the Combined dataset there are so many duplicates values in the movie titles we need to drop them.

Handling Missing Values:

We need to identify the missing values in the dataset.

We got that there are 6763 missing values in the dataset

Dropped all the columns which has more than 600 missing values =>Budget, USA\_gross\_income, Worldwide\_gross\_income and metascore.

Filled few columns with suitable values country with country unavailable, Cast with No Cast , Rating with No Rating , Description with no Description.

Dropped Avg\_votes columns.

Dropped columns which are repeated ,which we will not use for the predictions.

Data Cleaning:

For text data such as movie descriptions, cast, and genres, perform cleaning operations:

* Remove non-alphanumeric characters.
* Convert text to lowercase.
* Remove punctuations and stop words.

Utilize natural language processing techniques to clean and preprocess textual information.

Feature Engineering:

Create new features that could enhance model performance or provide additional insights:

Introduce features like 'votes\_class,' 'label' (IMDb rating classification), 'duration\_classification,' 'year\_classification,' etc.

*Preprocessing*: To streamline the analysis and modeling process, attributes that are not deemed primary for this project, such as date\_published usa\_gross\_income, worlwide\_gross\_income and metascore will be removed during the preprocessing stage. This focused selection ensures a more concise and relevant dataset, emphasizing key features that contribute significantly to the film recommendation and IMDb prediction objectives.

**Encoding:**

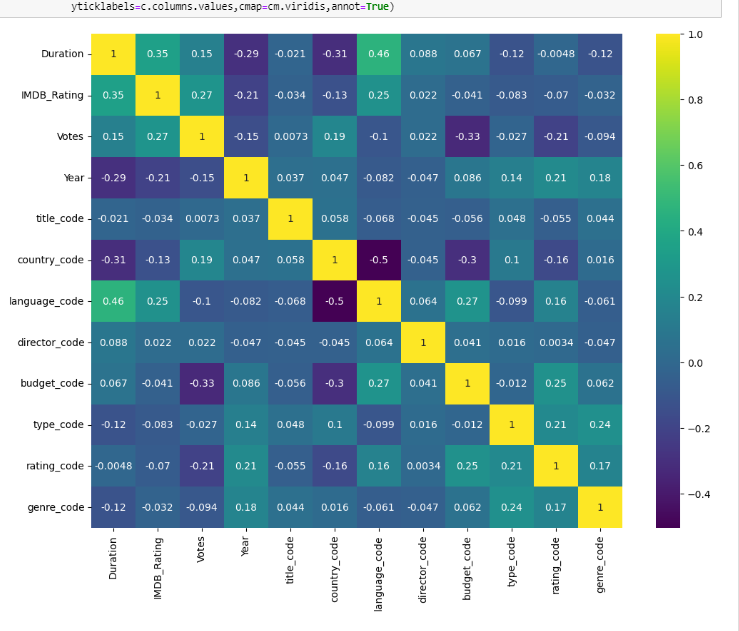
I performed encoding to convert the columns that have string values into columns that have integer values in order to use these columns in the algorithms that needs integer values in features.With the help of Sklearn's OrdinalEncoder I converted them into integer.

**Correlation:**

Now the data has all the numerical data we can plot how each attribute is connected with another.

For that we have considered pearson's and spearman's correlation.

Pearson’s:



Spearson’s:



**Models:**

**Unsupervised Models:**

We have used two Unsupervised Machine Learning Algorithms for Recommendation. We have used Kmeans and Cosine similarity Algorithms. First we have vectorized the left attributes in order to use them in the machine learning Models by removing alpha numeric characters, stop words , punctuations in the categorical attributes such as Genere, Description, Cast.

First we have used Kmeans algorithms in that we have used elbow method to the find the number of clusters and then used kmeans to pedict the clusters which represent the similar tv shows or movies. In this We have used PCA to reduce the dimensions. Overall we have observed that KMeans performed pretty well.

Using Kmeans we have clustered the similar movies together. So, it can easily recommend the movie based on the previously watched movie.

In cosine similarity, for each row based on their Title we put 10 columns (Duration, Country, Language, Director, Description, IMDb Rating, Cast, Year, Rating, Genre) into one column ("All Columns") to represent our bag of words.

Then created cosine similarity matrix and got result which was giving the recommendation.

**Supervised Models:**

Used Supervised Algorithms for IMDb Rating Prediction. First we have used Classification Models .In the classifications Models we have used Random Forest and KNN.

Random Forest

* 1. Used Random forest classifier To improve our algorithm's performance and increase the accuracy, we used different methods such as using hyperparameters and using some important features too.
  2. used GridSearchCV .
* KNN:
  1. First defined some new columns in order to get optimised result.
  2. Then did vectorization on string containing columns.
  3. Divided dataset into three parts - **train set, validation set, test set**
  4. Applied KNeighborsClassifier .

Then used for 4 Regression Models Linear Regression, KNN, Random Forest, Decision Tree regressors and measured the Mean Square Error, Mean Absolute Error, Root Mean Square Error.

**Implementation:**

Implemented total 8 models

2 unsupervised models

6 supervised models including both classification and regression below git hub link contains the implementation code.

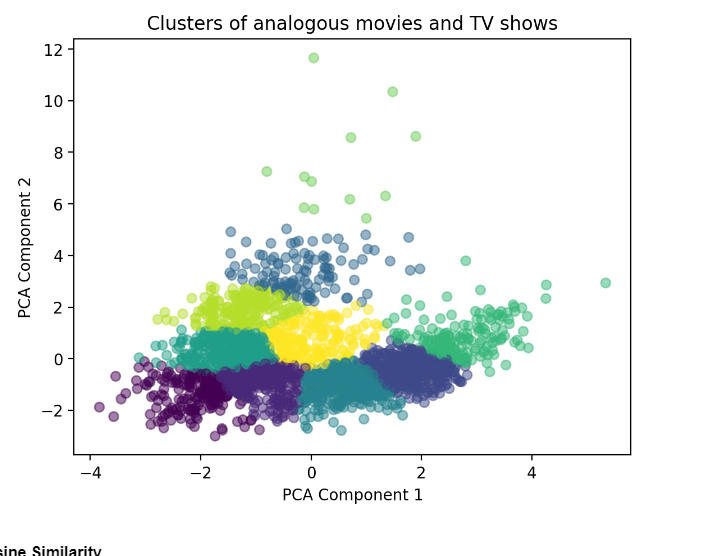
[**https://github.com/vaishnavi71/FilmRecommendationAndIMDBRatingPrediction**](https://github.com/vaishnavi71/FilmRecommendationAndIMDBRatingPrediction)

**Preliminary Results:**

**Recommendation System:**

**Kmeans:**

The Below Image shows the result of Kmeans which shows the clusters which represents the similar TV shows or movies together. So we can easily recommend a movie based on previously watched movie.



**Cosine Similarity:**

The below image shows the result of recommended movies based on the previously watch movie.

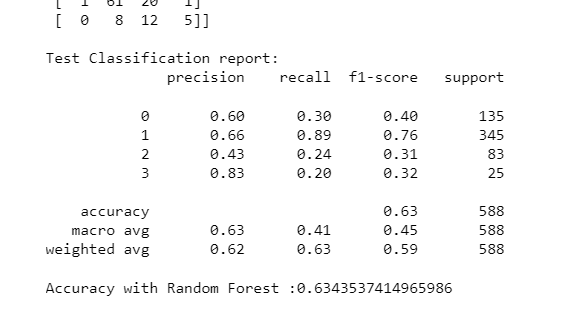
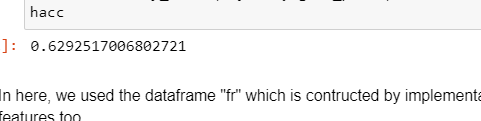


**IMDB Rating Prediction System:**

**Classification:**

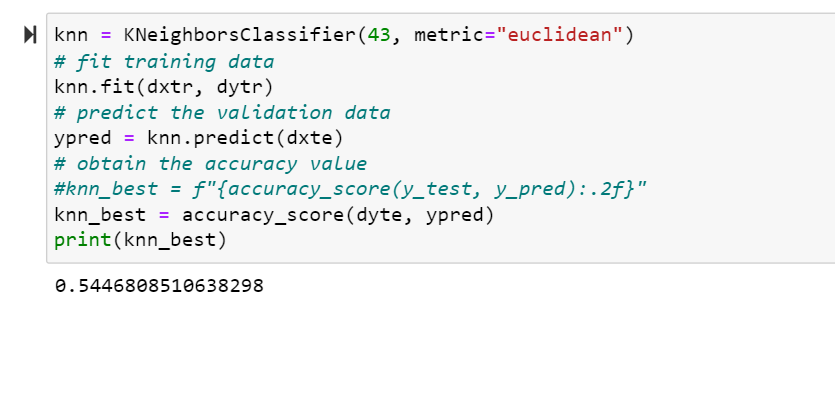
**Random Forest classifier:**

We get 63% accuracy with the random forest classifier and after vectorization , applying grid search cv we get 62% accuracy.

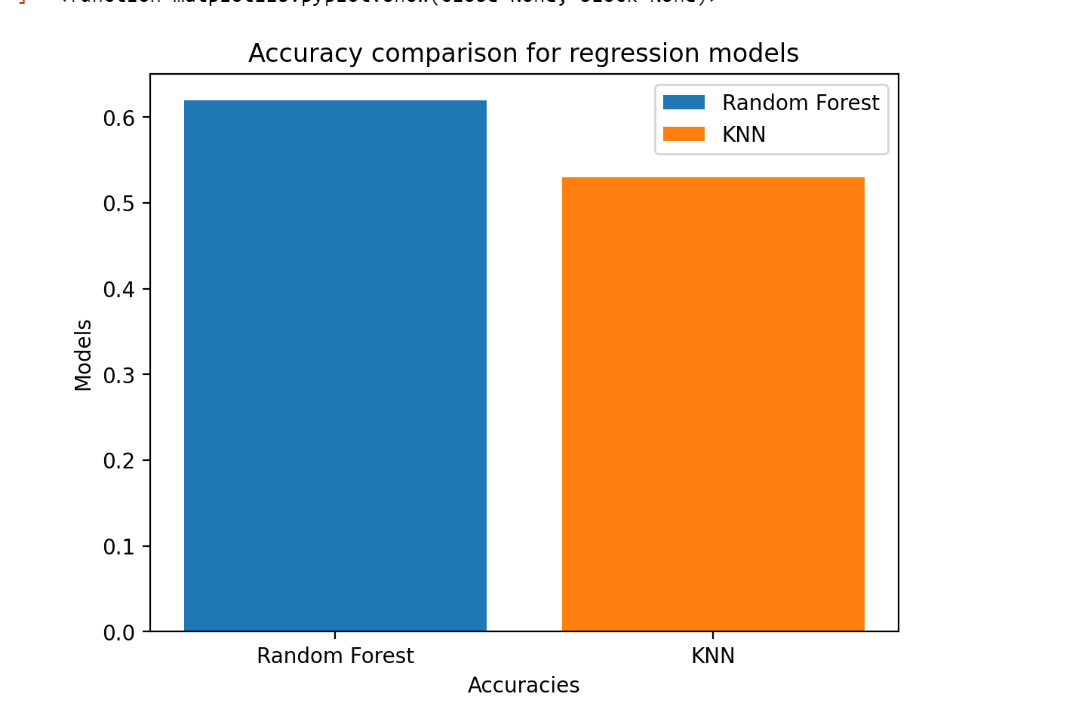
 

**KNN classifier:**

We get 54% Accuracy with the KNN classification even after vectorization.

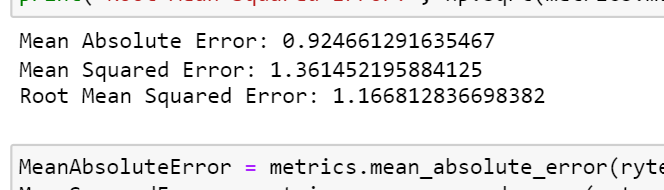


We have compared the 2 classifiers we can see that the Random Forest Classifier performs better compared to KNN Classifier.

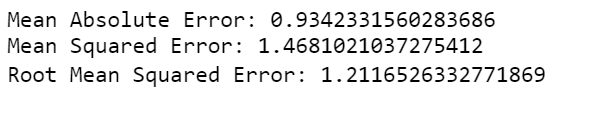


Then we have compared the mean square error, mean absolute error, root mean square of all the four regressors and we have observed that linear regression performs well when compared to other three regressors.

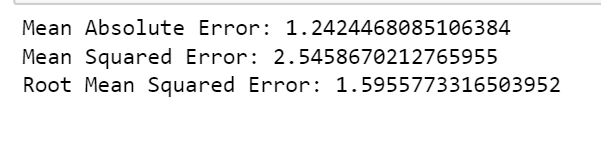
Linear Regression:



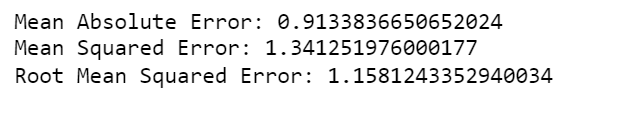
Random Forest Regression:

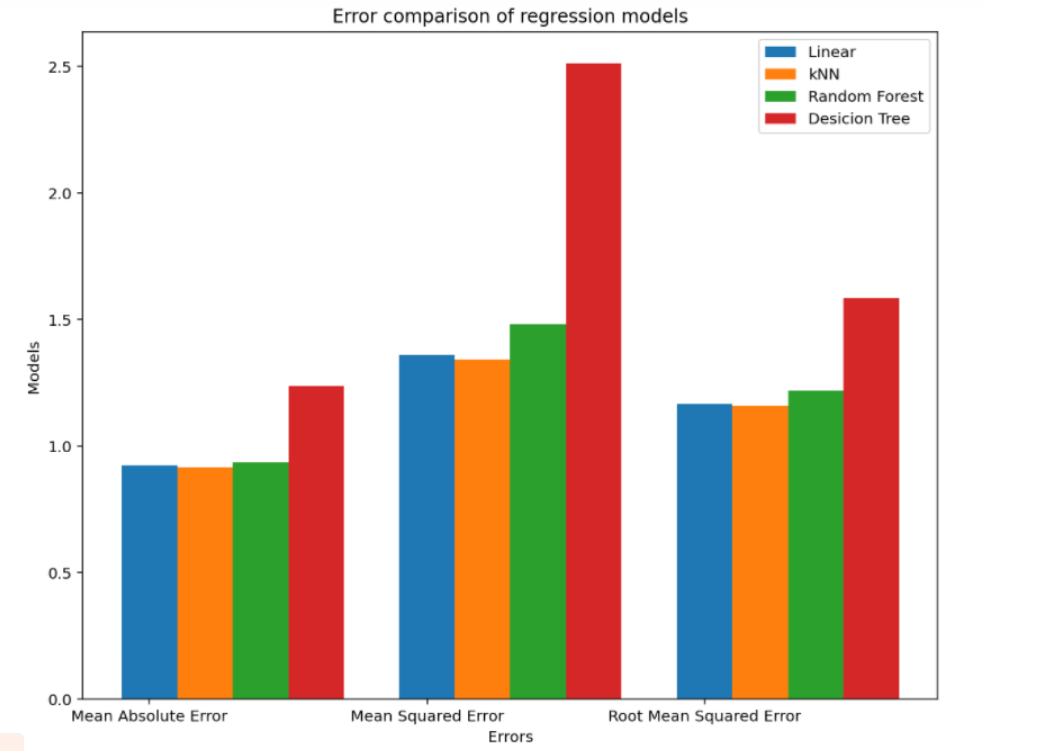


Decision Tree Regression:



KNN Regression:





**Project Management:**

***Work completed:***

1. Data Understanding and Exploration:
   * Task: Explore the dataset to understand its structure and contents.
   * Description: Perform preliminary data exploration to identify key features, data types, and potential challenges.
   * Responsibility (Task, Person): Data Analyst
   * Contributions (Members/Percentage): whole team
2. Data Cleaning:
   * Task: Cleanse the dataset by handling missing values, outliers, and inconsistencies.
   * Description: Implement data cleaning techniques to ensure the dataset is ready for analysis.
   * Responsibility (Task, Person): Data Science Team
   * Contributions (Members/Percentage): Akshaya Narsingu, Sai Chanikya Eagam
3. Correlation Analysis:
   * Task: Conduct correlation analysis on numerical features.
   * Description: Explore relationships between numerical variables using correlation matrices.
   * Responsibility (Task, Person): Data Analyst
   * Contributions (Members/Percentage): Sai Chanikya Eagam, Vaishavi Bommagani, Neelesh Bagal
4. Text Data Preprocessing:
   * Task: Preprocess textual data (Cast, Description, Genre) for feature extraction.
   * Description: Remove stop words, special characters, and perform tokenization on text features.
   * Responsibility (Task, Person): Data Science Team
   * Contributions (Members/Percentage): Akshaya Narsingu, Sai Chanikya Eagam, Vaishavi Bommagani
5. Feature Engineering:
   * Task: Create new features based on the insights from data exploration.
   * Description: Develop additional features that might enhance model performance.
   * Responsibility (Task, Person): Data Science Team
   * Contributions (Members/Percentage): Vaishavi Bommagani, Neelesh Bagal, Akshaya Narsingu
6. Machine Learning Model Development:
   * Task: Implement various machine learning models for IMDb rating prediction.
   * Description: Develop regression models considering features like Votes, Duration, Budget, etc.
   * Responsibility (Task, Person): Machine Learning Engineer
   * Contributions (Members/Percentage): Vaishavi Bommagani, Neelesh Bagal
7. Model Evaluation - Regression:
   * Task: Evaluate regression models using metrics like Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error.
   * Description: Assess the accuracy and reliability of regression models.
   * Responsibility (Task, Person): Data Analyst
   * Contributions (Members/Percentage): Vaishavi Bommagani

***Work to be completed:***

1. Model Optimization:
   * Task: Fine-tune hyperparameters to optimize model performance.
   * Description: Use techniques like Grid Search to find the best parameters for the models.
   * Responsibility (Task, Person): Machine Learning Engineer
   * Contributions (Members/Percentage): TBD
2. Implementation Status Report - Preliminary Results:
   * Task: Generate an implementation status report for preliminary results.
   * Description: Summarize the progress, findings, and challenges encountered during the initial phase.
   * Responsibility (Task, Person): Project Manager
   * Contributions (Members/Percentage): TBD
3. Documentation and Reporting:
   * Task: Document the entire project, including methodologies, results, and recommendations.
   * Description: Create comprehensive documentation for future reference and reporting.
   * Responsibility (Task, Person): Project Manager
   * Contributions (Members/Percentage): TBD

**References/Bibliography:**

1.Comprehensive Movie Recommendation system **,**[HrisavBhowmick](https://arxiv.org/search/cs?searchtype=author&query=Bhowmick,+H), [Ananda Chatterjee](https://arxiv.org/search/cs?searchtype=author&query=Chatterjee,+A), [Jaydip Sen](https://arxiv.org/search/cs?searchtype=author&query=Sen,+J),

[**https://doi.org/10.48550/arXiv.2112.12463**](https://doi.org/10.48550/arXiv.2112.12463)

**2.**Review of Movie Recommendation system**,** [**https://ieeexplore.ieee.org/document/9785014**](https://ieeexplore.ieee.org/document/9785014)

**3.** [**https://github.com/kishan0725/AJAX-Movie-Recommendation-System-with-Sentiment-Analysis/tree/master**](https://github.com/kishan0725/AJAX-Movie-Recommendation-System-with-Sentiment-Analysis/tree/master)

**Conclusion:**

Taking everything into consideration, For IMDb prediction algorithms, in clasification part, at first both KNN and RF have close accuracies but after implementing the conter vectorizer, we saw that RF gives a higher accuracy value. For the regression, we concluded that linear regression was performed better when we compare it with KNN, RF and Decision Tree regressions. Also, for our recommendation systems cosine similarity performed better than k-means.

For our recommendation systems, cosine similarity's results are more consistent.