FINAL PROJECT REPORT

Course Name: DATA*6300 ANALYSIS OF BIG DATA

Project Name: NLP-driven Content based Movie Recommendation, Genre Classification and Movie Ratings Prediction.

Name: Mohammad Sakibul Islam
ID: 1265299

1.0 - Introduction:

With the plethora of movies available on the internet these days, viewers often face difficulty in finding the movie that aligns with their interests. Relying entirely on genre-based searches may overlook the individual preferences of the viewer, which actually emphasize the importance of content based movie suggestions. By leveraging movie descriptions and plot summaries, content based movie recommendation systems can facilitate the exploration of similar movies that match with the viewer preferences.

Moreover, genre classification based on movie plot descriptions facilitates the analysis of movie content for efficient genre categorization of movies as it enables the identification of the most frequently used words in movie description or relevant movie themes associated with genres. In addition, accurate genre classification is useful for audiences to find out the movies that align with their preferred genres.

Furthermore, understanding audience preferences and gaining insights into the factors influencing movie success is crucial in the film industry, which highlights the need for exploratory analysis with movie ratings prediction. Since, analysis of movie ratings involves taking audience satisfaction factors into account, it offers insights into the key features that contribute to the movie's success and helps us to understand audience preferences.

The motivation behind this project lies in exploring similar movies for audiences, analyzing movie descriptions to determine the underlying theme of the movie and highlighting the most important factors that impact the audience satisfaction and movie ratings. So, this project is significant as it aims to enhance movie exploration and improve user satisfaction by suggesting more relevant movies, and effectively categorizing movies into genres based on movie content. Additionally, the project holds importance because it seeks to uncover the key features which are associated with audience engagement.

2.0 - Problem Statement:

The primary objective of the project is to employ natural language processing (NLP) and machine learning methods to derive meaningful insights from movie dataset. The project comprises three tasks:

- **Content-based Movie Recommendation:** The project aims to leverage NLP techniques to develop a content-based movie recommendation system which will suggest movies based on the similarities in movie descriptions. In order to suggest more relevant movies that align with viewer choice, the project focuses on analyzing the movie descriptions text data and identifying the semantic similarities between the movies.

- **Genre Classification based on Movie content:** The project seeks to employ NLP techniques to analyze the textual content from the movie overviews and develop classification models to categorize movies into multiple genres simultaneously. So, this task is a multilabel classification problem, since a movie can belong to multiple genres at once.

- **Movie Ratings Prediction:** By analyzing various movie features such as genre, budget, revenue, runtime, language and production companies, the project aims to predict movie ratings by utilizing machine learning techniques. Moreover, by conducting feature importance analysis, the project seeks to identify the important features impacting the target movie rating variable.

So, the purpose of this project is to enhance the movie watching experience of the audiences by conducting comprehensive analysis of the movie dataset and suggesting similar movies and effective content categorization utilizing NLP along with highlighting the critical factors which directly influence audience satisfaction and engagement for movies.

3.0 - Method:

3.1 - Dataset Information and Preprocessing:
Through web scraping from IMDb website, data on 50 movies with 26 feature columns have been extracted initially. Then, considering the time consuming nature and potential constraints of extracting large amounts of data, I have decided to merge the web scraped data with a publicly available movie dataset[1] from kaggle, which contains 1008166 movies with 23 features. After merging both the dataset obtained from kaggle and webscraping, and conducting preprocessing steps, my final dataframe comprises 432410 movies with 20 feature columns. The dataframe contains various movie features including title, runtime, average rating, rating count, popularity, budget, revenue, language, production companies, genres, overview, and release date.

Data Preprocessing Steps:

1. **Handling Duplicate Movie Names:** Duplicate movies are identified and removed to ensure each movie is represented uniquely.
2. **Handling Missing Values:** Columns with high percentage of missing values are removed and missing values for some categorical columns are imputed with suitable values.
3. **Data Type Conversion:** For ensuring consistency in the analysis, “release_date” column has also been converted from string to date type. Moreover, columns such as “budget”, “revenue” and “avg_rating” are converted to double type and “runtime” and “rating_count” are changed to integer type.
4. **Handling Outliers:** Boxplot is used to visualize the outliers for movie runtime. Movies with unusual runtimes are removed as they significantly differ from most other observations.
5. **Feature Extraction:** Both the release year and release month are extracted from the release date for conducting temporal analysis.
6. **One-hot Encoding for Genres:** One-hot encoding is applied to categorical genres column to convert it into binary features for each genre category.
7. **Text Preprocessing:** Some preprocessing techniques including removing punctuations, eliminating stop words (commonly used words that do not carry significant meaning) and

converting text to lowercase have been performed on movie overviews to prepare text data for genre classification and movie recommendation.

8. **Removing Unnecessary columns:** For movie ratings prediction, some columns which are irrelevant to the target variable are removed.
9. **Scaling numerical features:** For movie ratings prediction, scaling of numerical features is conducted in order to prevent feature dominance and ensure stability and uniformity.

3.2 - Modelling:

(i) Movie Recommendation System:

The project explores two different approaches for content-based movie recommendation: Word2Vec approach and TF-IDF approach.

(1) Word2Vec Approach: Word2Vec is a word embedding technique in Natural Language Processing (NLP) which captures the semantic relationship between words. For effective content based analysis, firstly, “title”,”overview” and “genre” columns are concatenated to create the ‘content’ column. Then, after applying some preprocessing techniques (removal of punctuation and stop words, conversion to lowercase and tokenization of text into words) on the “content” column, word2vec model is trained using the tokenized data and subsequently, embeddings for each movie are calculated by averaging the word embeddings of each movie’s words. Finally, after calculating the cosine similarity between the embedding vectors of the selected movies and all other movies, top 10 similar movies are recommended based on the similarity scores.

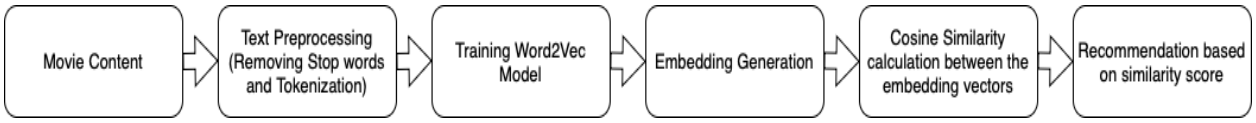


Fig: Overview of Word2Vec approach

(2) TF-IDF Approach: TF-IDF (Term Frequency-Inverse Document Frequency) is used in NLP (Natural Language Processing) to determine how important a term is to a document in a collection of documents. So, TF-IDF combines two factors: Term Frequency (TF) which reflects how frequently a word appears in a document and Inverse Document Frequency (IDF) which reflects how unique a word is across all the documents.

$$\text{TF} = \frac{Number\ of\ times\ the\ specific\ term\ appears\ in\ a\ document}{Total\ Number\ of\ terms\ in\ a\ document}$$

$$\text{IDF} = log(\frac{Number\ of\ documents}{Number\ of\ documents\ that\ contains\ the\ \ specific\ term})$$

$$\text{TF-IDF} = \text{TF} * \text{IDF}$$

In this approach, firstly movie overviews, titles and genres are combined into movie content, and then after applying some preprocessing to text data, TF-IDF vectorization is used to convert it

into numerical vectors. Finally, cosine similarity is calculated between the TF-IDF vectors, and top 10 similar movies are suggested based on the similarity scores.

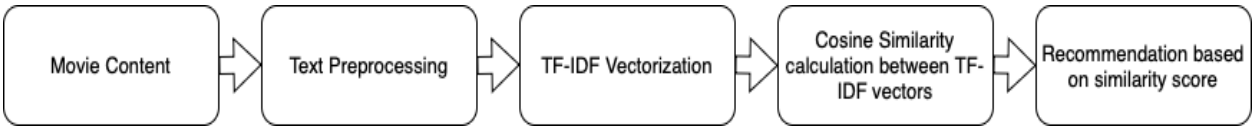


Fig: Overview of TF-IDF approach

(ii) Movie Genre Classification:

Movie Genre Classification is a multilabel classification problem because a movie can be associated with multiple genre classes simultaneously. For classification, some preprocessing techniques are applied to movie content text data, including converting to lowercase, removing stop words and performing lemmatization (reducing a word to base form). The processed text is then converted into numerical vectors using TF-IDF vectorization.

The dataset is split into training and test sets, with 80% data allocated for training and 20% for testing. For multilabel movie genre classification, various classification models, such as Logistic Regression, Random Forest, Gradient Boosting and Adaboost Classifier are trained. Logistic regression model is chosen for its simplicity and other ensemble learning methods (Random Forest, Gradient Boosting and Adaboost Classifier) are selected because these models combine predictions from multiple base weak learners and are less prone to overfitting.

As a movie can belong to more than one genre class simultaneously, Multi-output classifiers are used to manage the prediction of multiple genre classes. Finally, model performances are measured using precision, recall and f1-score as evaluation metrics.

Fig: workflow of content based genre classification

(iii) Movie Ratings Prediction:

After applying some preprocessing steps and removing the irrelevant feature columns, the dataset is divided into training and test set, where 80% data used for model training and 20% data allocated for evaluation. Various regression models such as Linear Regression, Ridge Regression, Support Vector Regressor, Decision Tree Regressor, Random Forest, Extreme Gradient Boosting and Bagging Regressor are trained to capture the relationship between the input features and target movie rating variable.

Linear Regression is selected for simplicity, while Ridge regression is chosen for handling multicollinearity. To effectively capture complex non-linear relationships, Support Vector Regressor and Decision Tree Regressor are chosen for ratings prediction. Ensemble learning

methods such as Random Forest, Extreme Gradient Boosting and Bagging Regressor are selected because these models aggregate the predictions of multiple base learners which helps to reduce overfitting and improve generalization.

During model training for movie ratings prediction, k-fold cross validation is employed to reduce the risk of overfitting, where data is split into k folds and the model is trained k times. Each time k-1 folds are used for training and the remaining one fold is used for validation. Finally, the performance of the trained models is measured and compared using mean squared error (MSE) and mean absolute error (MAE) as evaluation metrics.

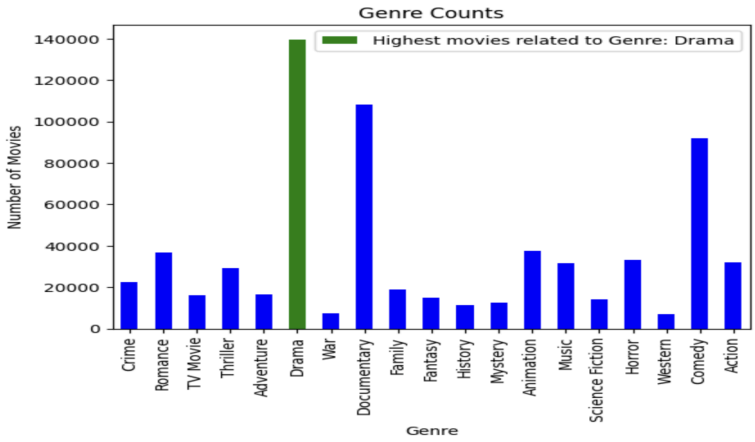
4.0 - Results and Discussion:

4.1 - Results for Exploratory Data Analysis:

Exploratory data analysis (EDA) is performed on various movie features, such as genres, runtime, budget, revenue, release date, average rating and popularity to uncover the trends, patterns and relationship among variables, and gain meaningful insights for the project's objectives.

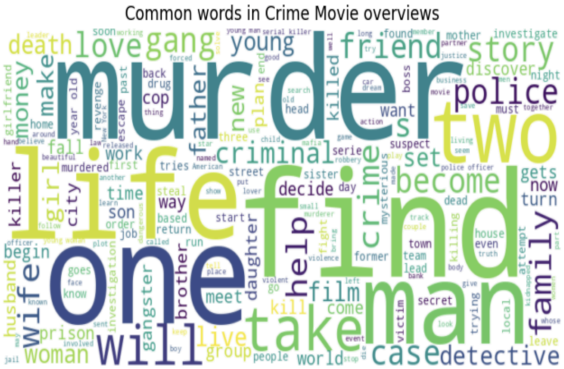
Genre Analysis:

(i) Question: How does the distribution of movie counts vary across different genres in the dataset?



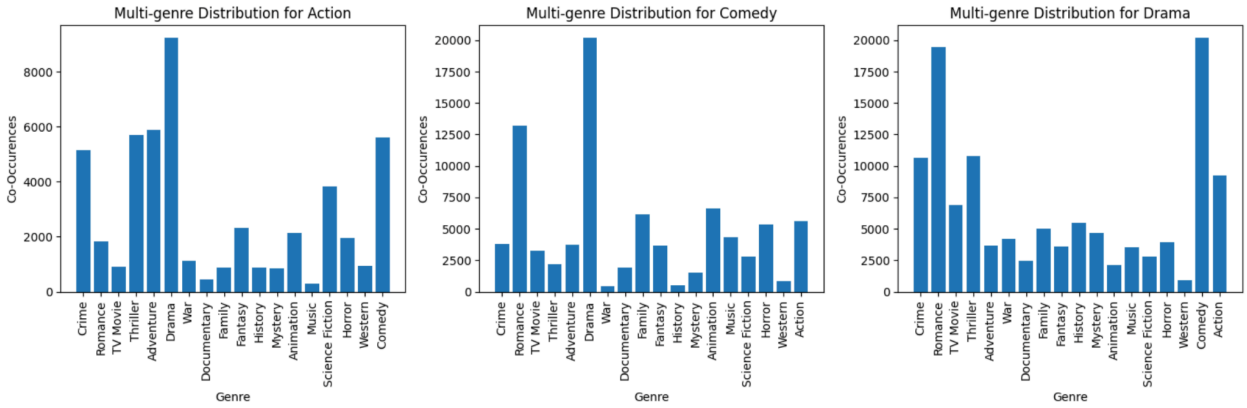
Insight: From the visualization, it is evident that the highest number of movies belong to the drama genre, followed by documentary and comedy in the dataset.

(ii)Question: What are the most common words in the movie overviews across different genres?



Insight: The word cloud which represents the most frequently used words in movie description of each genre. For instance, the most commonly used words in crime genre movie content include “murder”, “police”, “life”, whereas for romance genre movies, the most prominent words used are “love”, “life”, “find” and “fall”. These insights are really useful for understanding the content and plot variations across different genres.

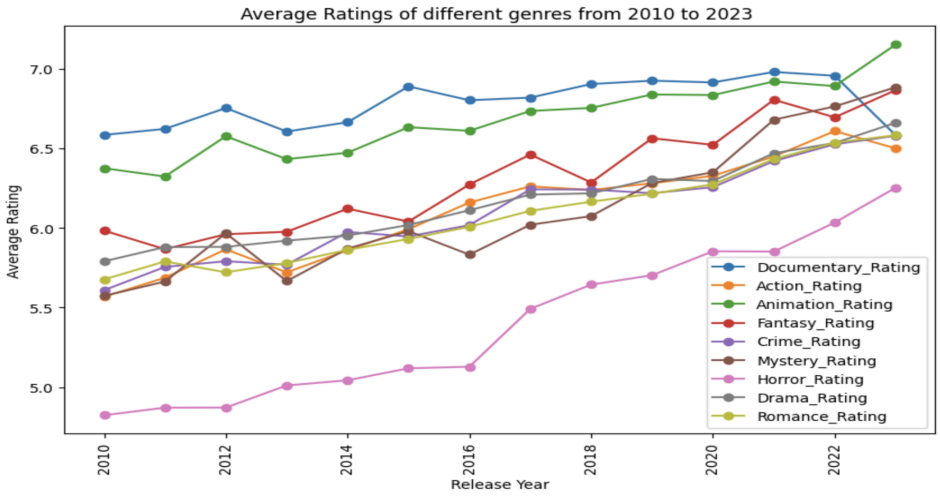
(iiii) Question: What are the distributions of co-occurrences of other genres when a movie is specified as action, comedy or drama?



Insight: The visualization shows that Action genre movies frequently co-occur with drama, adventure and thriller genre. Moreover, Comedy genre movies overlap highly with drama genre and drama genre movies commonly co-occur with comedy and romance genre. So, this analysis provides insights into potential patterns in content categorization by visualizing the co-occurence of movie genres.

Rating Analysis:

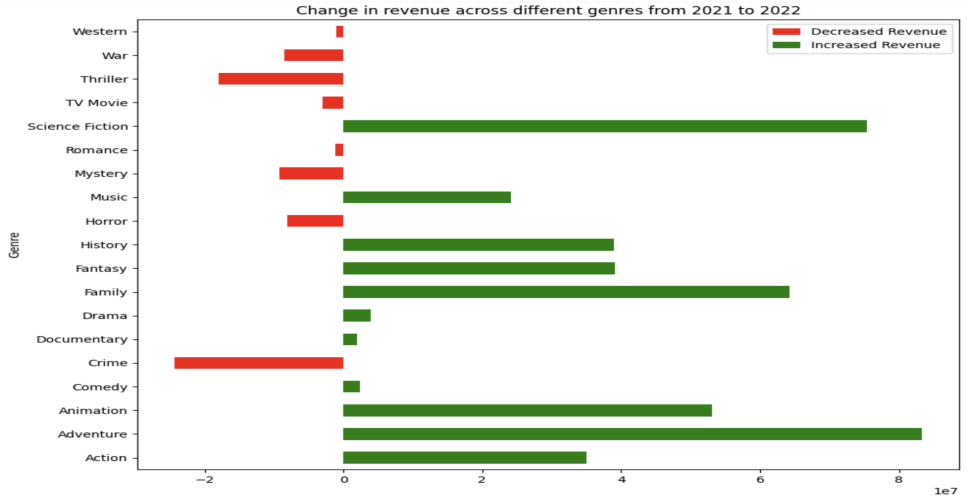
(i) Question: How does the average movie rating across different movie genres change over a period from 2010 to 2023?



Insight: Documentary and animation genre movies receive higher average ratings compared to the other movie genres from 2010 to 2023. Although there are some fluctuations, the average ratings of each genre show overall increase over the period. This analysis provides insights about the trends of audience preference and interests across different genres over the years.

Revenue Analysis:

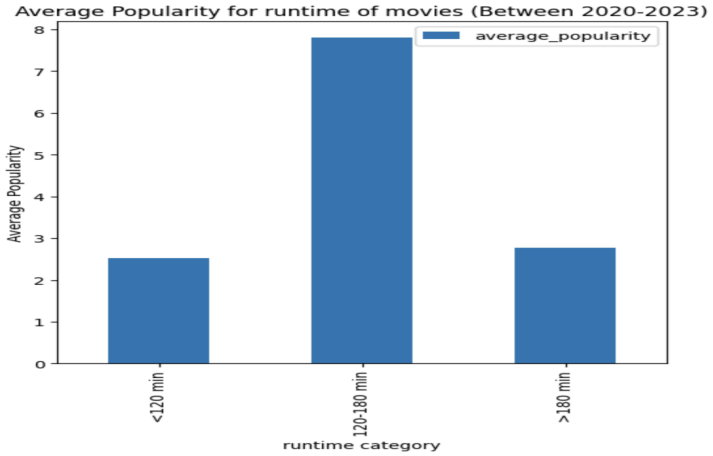
(i) Question: How has the average movie revenue changed between 2021 and 2022 across different movie genres?



Insight: The average revenue decreased for crime, horror, mystery, romance, tv movie, war and western genre movies from 2021 to 2022, whereas there is an increase in the revenue in the other genres including action, adventure, animation and comedy.

Runtime Analysis:

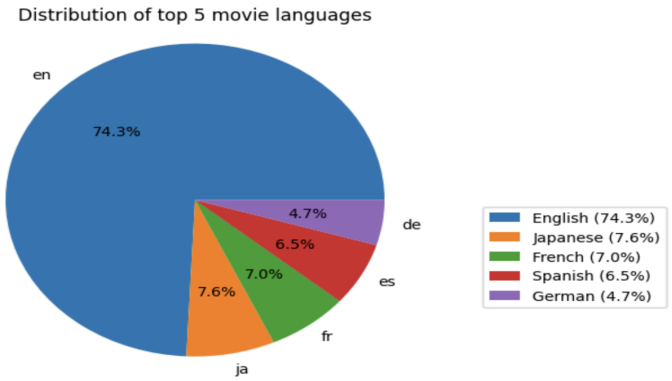
(i) Question: How does the runtime of movies correlate with movie popularity metrics?



Insight: The visualization represents that movies having runtime between 120-180 minutes have the highest average popularity compared to the other category of runtime, which indicates that audiences mostly prefer movies having a moderate runtime between 120 to 180 minutes.

Movie Language Analysis:

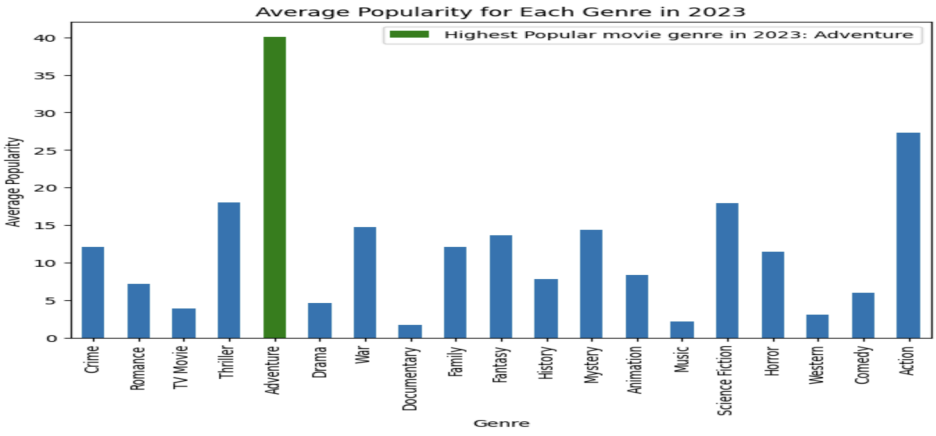
(i) Question: What is the distribution of movie languages in the dataset? Which language is most prevalent in terms of frequency?



Insight: The pie chart shows that a higher proportion of movies are in English language, which shows a dominance of English language movies in the dataset.

Popularity Analysis:

(i) Question: Which movie genres are most popular in the year 2023?

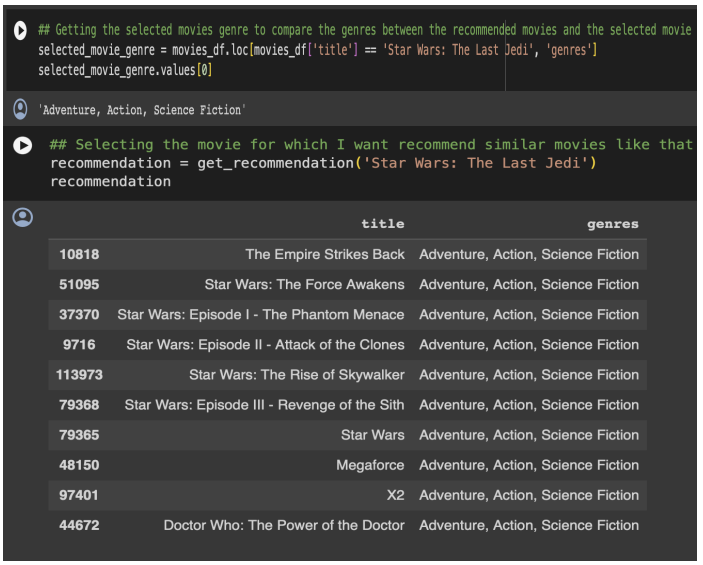


Insight: In the year 2023, the most popular movie genre is adventure, followed by action and science fiction. This analysis provides insights about the popularity of various genres among the viewers.

4.2 - Machine Learning Results:

(i) Movie Recommendation System Results:

Word2Vec approach for content based movie recommendations yields effective results, since the suggested movies are similar and belong to the same genres as the chosen movie. The attached image shows the recommended movies for ‘Star Wars: The Last Jedi’ using the Word2Vec approach which demonstrates that both the selected movie and the recommended movies belong to the same genre. Also, the suggested movies such as ‘Star Wars: The Force Awakens’, ‘Star Wars’ are similar and closely related to the selected movie, which indicates the effectiveness of this approach.



```
## Getting the selected movies genre to compare the genres between the recommended movies and the selected movie
selected_movie_genre = movies_df.loc[movies_df['title'] == 'Star Wars: The Last Jedi', 'genres']
selected_movie_genre.values[0]
```

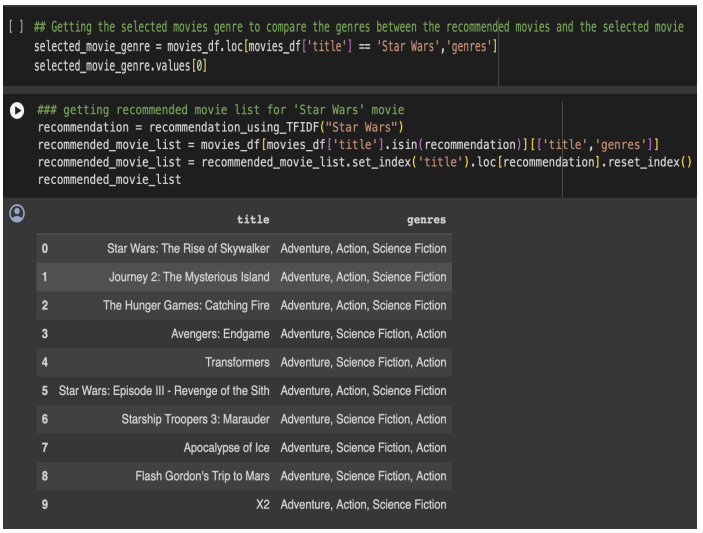
```
'Adventure, Action, Science Fiction'
```

```
## Selecting the movie for which I want recommend similar movies like that
recommendation = get_recommendation('Star Wars: The Last Jedi')
recommendation
```

	title	genres
10818	The Empire Strikes Back	Adventure, Action, Science Fiction
51095	Star Wars: The Force Awakens	Adventure, Action, Science Fiction
37370	Star Wars: Episode I - The Phantom Menace	Adventure, Action, Science Fiction
9716	Star Wars: Episode II - Attack of the Clones	Adventure, Action, Science Fiction
113973	Star Wars: The Rise of Skywalker	Adventure, Action, Science Fiction
79368	Star Wars: Episode III - Revenge of the Sith	Adventure, Action, Science Fiction
79365	Star Wars	Adventure, Action, Science Fiction
48150	Megaforce	Adventure, Action, Science Fiction
97401	X2	Adventure, Action, Science Fiction
44672	Doctor Who: The Power of the Doctor	Adventure, Action, Science Fiction

Fig: Recommended movies using Word2Vec

TF-IDF approach for movie recommendation also proves to be effective. The attached image demonstrates the suggested movies for ‘Star Wars’ movie using TF-IDF approach and the recommended movies for this selected movie includes ‘Star Wars: The Rise of Skywalker’ which perfectly aligns with the genres and content of the selected movie. Both the selected movie and the recommended movies belong to ‘Adventure’, ‘Action’ and ‘Science Fiction’ genres. This represents that the TF-IDF approach effectively suggests movies with similar characteristics.



```
[ ] ## Getting the selected movies genre to compare the genres between the recommended movies and the selected movie
selected_movie_genre = movies_df.loc[movies_df['title'] == 'Star Wars','genres']
selected_movie_genre.values[0]
```

```
### getting recommended movie list for 'Star Wars' movie
recommendation = recommendation_using_TFIDF("Star Wars")
recommended_movie_list = movies_df[movies_df['title'].isin(recommendation)][['title','genres']]
recommended_movie_list = recommended_movie_list.set_index('title').loc[recommendation].reset_index()
recommended_movie_list
```

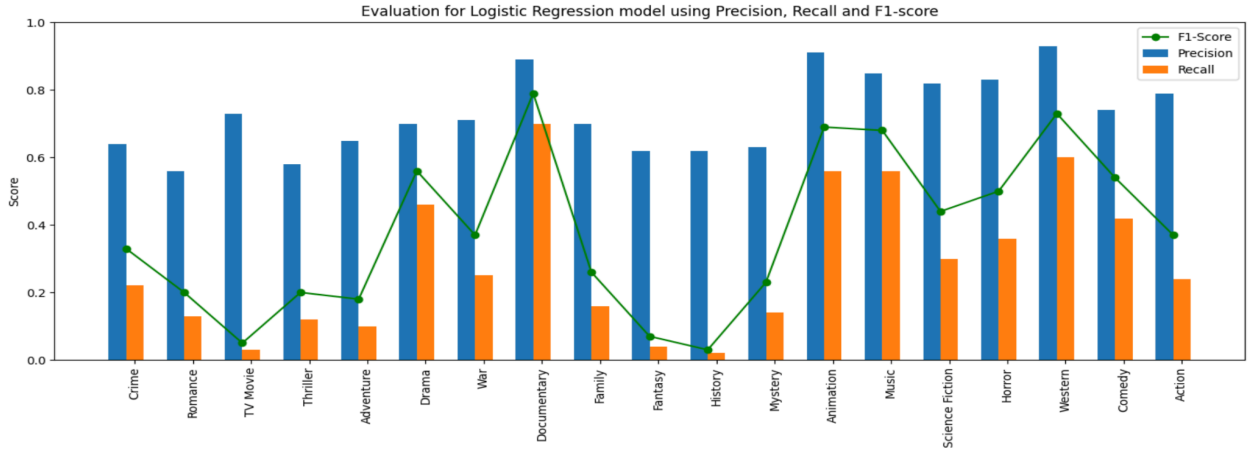
	title	genres
0	Star Wars: The Rise of Skywalker	Adventure, Action, Science Fiction
1	Journey 2: The Mysterious Island	Adventure, Action, Science Fiction
2	The Hunger Games: Catching Fire	Adventure, Action, Science Fiction
3	Avengers: Endgame	Adventure, Science Fiction, Action
4	Transformers	Adventure, Science Fiction, Action
5	Star Wars: Episode III - Revenge of the Sith	Adventure, Action, Science Fiction
6	Starship Troopers 3: Marauder	Adventure, Science Fiction, Action
7	Apocalypse of Ice	Adventure, Science Fiction, Action
8	Flash Gordon's Trip to Mars	Adventure, Science Fiction, Action
9	X2	Adventure, Action, Science Fiction

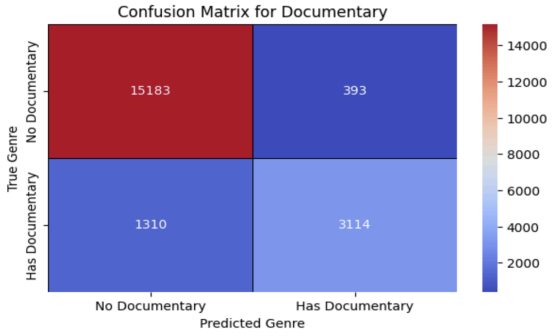
Fig: Recommended movies using TF-IDF approach

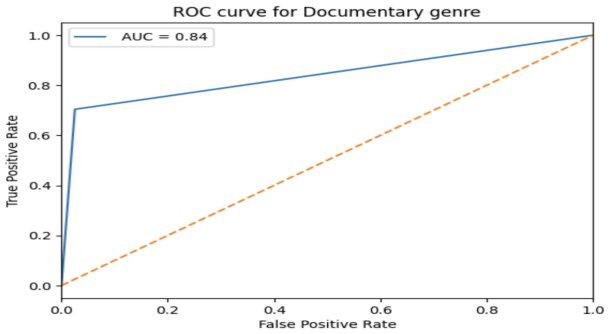
(ii) Genre Classification Results:

For multilabel genre classification, various classification models are evaluated based on precision, recall and f1-score to compare their performance.

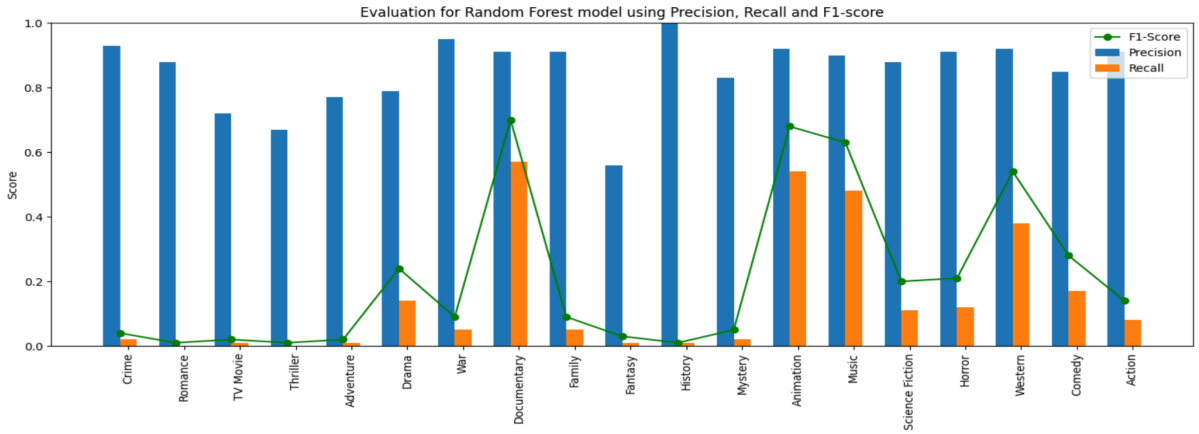
Logistic Regression: From the visualization of precision, recall and f1-score across different genres using logistic regression, it is evident that logistic regression model performs well in predicting the documentary, music, western and animation genres, but the model performs poorly in classifying family, fantasy, history and mystery genres as reflected by the low recall and f1-scores. However, the confusion matrix plot and ROC curve show that logistic regression model performs well in predicting the documentary genre because a higher AUC score(0.84) represents a good performance. The higher AUC score for documentary genre highlights that logistic regression model can effectively differentiate between documentary and other genres.

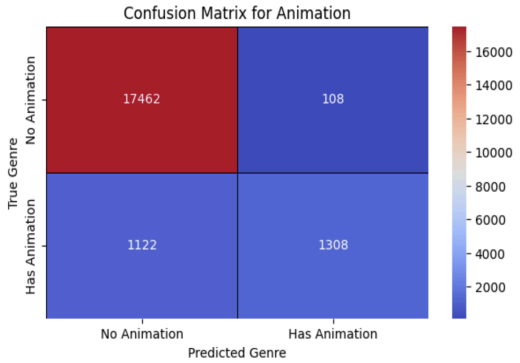


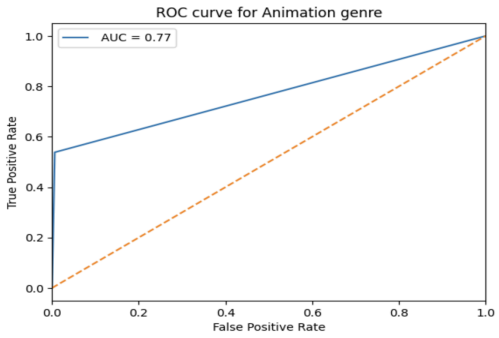




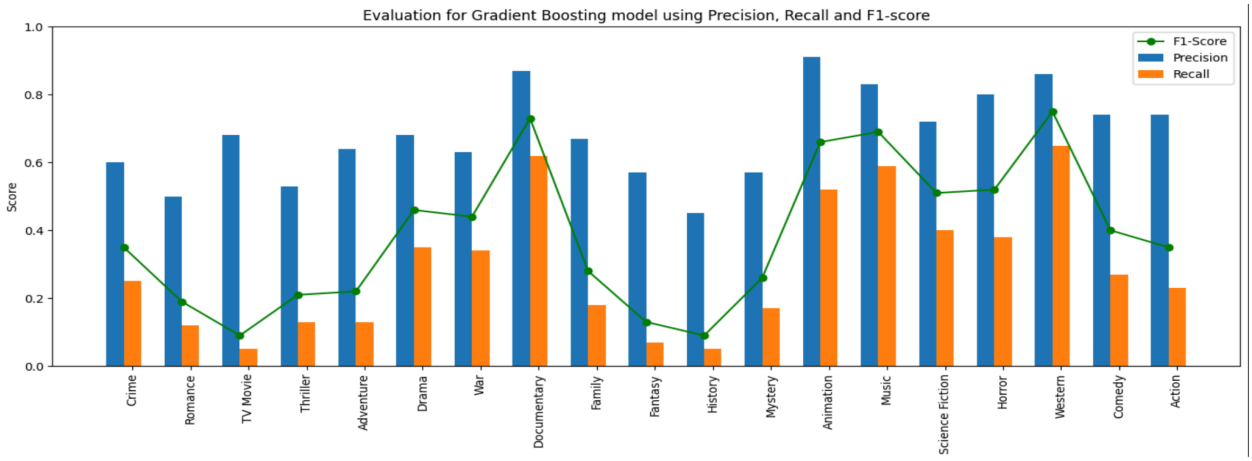
Random Forest: Random Forest model performs well in predicting the documentary, music, western and animation genres, but the model struggles to predict accurately for crime, romance, family, fantasy, history and mystery genres which is indicated by the low recall and f1-scores in the visualization. However, the random forest model performs well in predicting the animation genre which is demonstrated by the confusion matrix and the higher AUC (area under curve) in the ROC curve plot. The AUC for the animation genre is 0.77 using the random forest model which indicates that model is performing well in distinguishing between the animation and other genres.

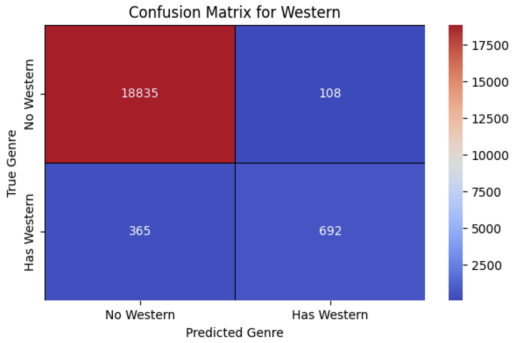


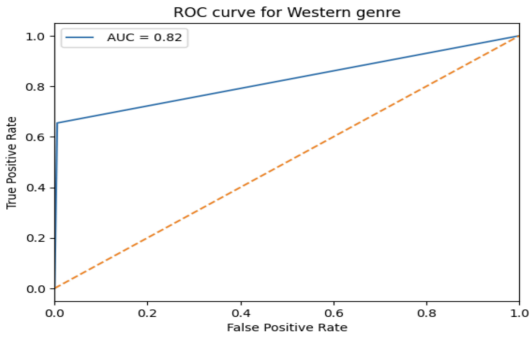




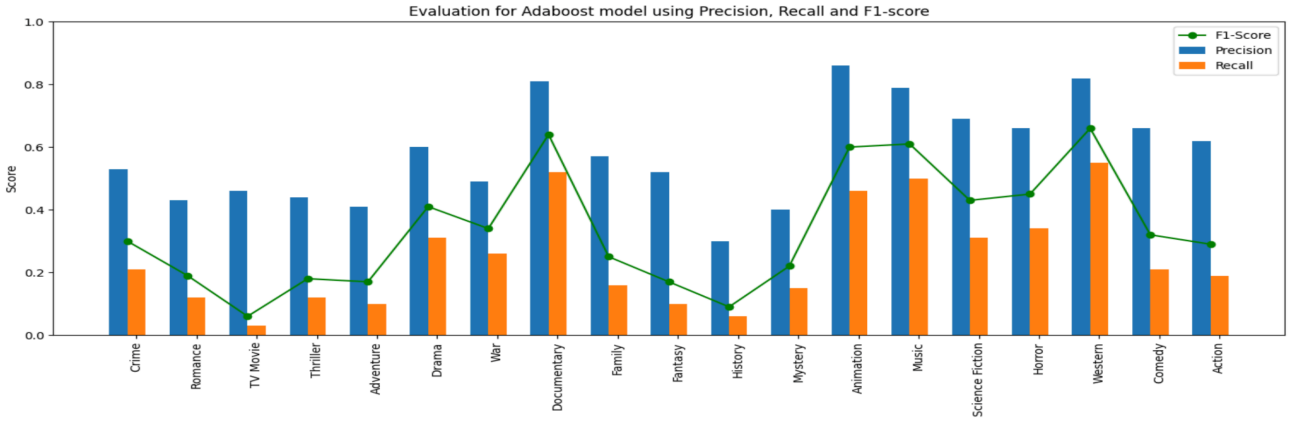
Gradient Boosting: Gradient Boosting model also performs well in predicting the documentary, music, western and animation genres, but the model demonstrates poor performance for tv movie, fantasy, history genres classification which is represented by the low recall and f1-scores in the visualization graph. However, the gradient boosting model performs notably well in predicting the western genre which is illustrated by the confusion matrix and the ROC curve plot.

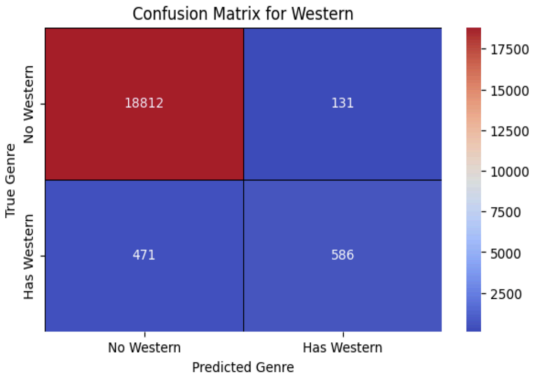


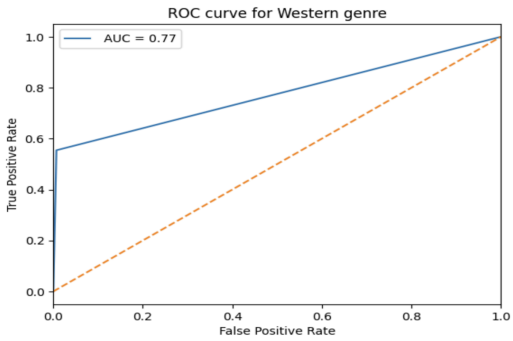




Adaboost Classifier: The visualization graph of precision, recall and f1-score shows that the adaboost classifier model has a decent performance in predicting the documentary, music, western and animation genres, but the model does not perform well while classifying for tv movie, fantasy, and history genres which is indicated by the low recall and f1-scores. However, the confusion matrix and the ROC curve plot represent that the adaboost classifier model performs moderately well in predicting the western genre, although there are some misclassifications. The higher area under the curve (AUC =0.77) for western genre illustrates that the model is effective for differentiating the western genre from other genres.

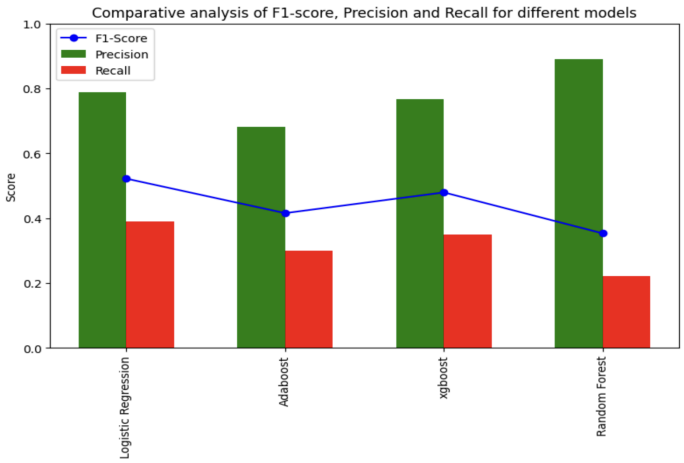






Comparative Analysis of Classification Models:
By comparing the precision, recall and f1-score of different models for multilabel genre classification based on content, it is evident that Logistic Regression model provides the better performance among all the considered classification models and the average F1-score for logistic regression model is 0.52.

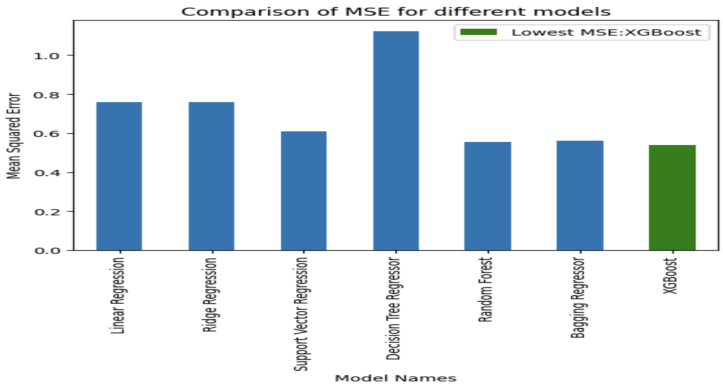
The average f1-score for Random Forest, Gradient boosting and Adaboost classifiers are 0.352, 0.479 and 0.415 respectively, which indicates that these ensemble models perform moderately in classification. One of the reasons behind lower f1-scores is the presence of imbalance class distribution and balancing the class is challenging as a movie can belong to multiple genres. Another reason is increased complexity to the classification problem due to prediction of multiple genre labels for each movie across 19 different genre classes. Moreover, misclassification of genre can occur due to overlapping characteristics between different genre classes, such as a movie with description elements of both Action and war genre may be difficult to classify accurately.



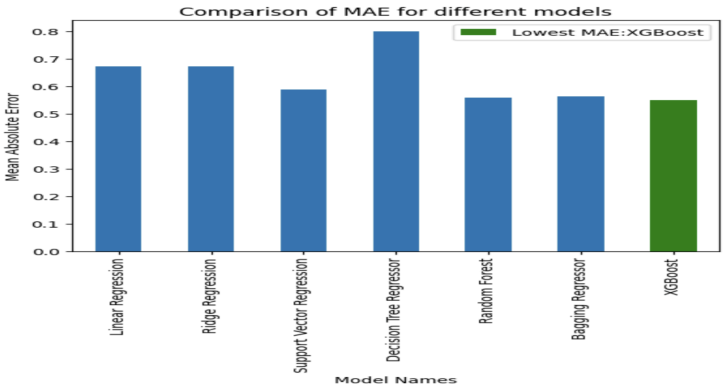
Model	Average Precision	Average Recall	Average F1-Score
Logistic Regression	0.788159	0.389746	0.521573
Adaboost	0.681506	0.298630	0.415285
xgboost	0.766295	0.348519	0.479127
Random Forest	0.888751	0.220140	0.352875

(iii) Movie Ratings Prediction Results:
In the movie ratings prediction task, performance of various regression models are measured using mean squared error(MSE) and mean absolute error (MAE). By comparing the MSE and MAE, it is evident that ensemble methods like Random Forest, Gradient Boosting and Bagging Regressor model outperforms the individual models (Linear Regression, Ridge Regression, Support Vector Regression and Decision Tree Regression) because ensemble methods combine predictions of multiple base models which mitigates the risk of overfitting.

As the gradient boosting model can capture the complex relationship between variables and optimize model performance by iteratively minimizing the loss function, it exhibits the better predictions having the lowest MSE and MAE among all the considered regression models. On the other hand, Decision Tree Regressor model provides the worst performance in predicting movie ratings as it is more prone to overfitting.

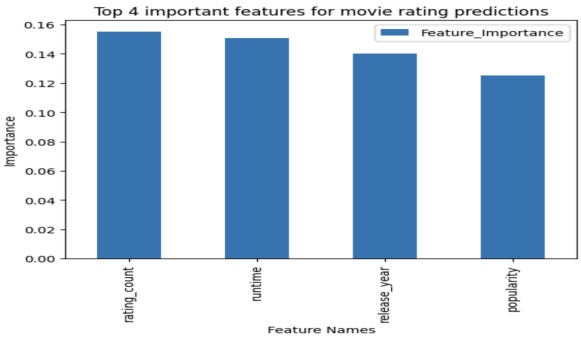


Model Name	MSE
Linear Regression	0.759056
Ridge Regression	0.759056
Support Vector Regression	0.608465
Decision Tree Regressor	1.122414
Random Forest	0.554258
Bagging Regressor	0.560916
XGBoost	0.540013



	Model Name	MAE
0	Linear Regression	0.671960
1	Ridge Regression	0.671960
2	Support Vector Regression	0.587701
3	Decision Tree Regressor	0.799995
4	Random Forest	0.559634
5	Bagging Regressor	0.563687
6	XGBoost	0.550314

Moreover, after conducting feature importance analysis, it is found that rating count, runtime, release year and popularity are the most influential factors impacting the target movie rating variable.



5.0 - Conclusion:

The project is focused on employing NLP and machine learning techniques to address three challenges: providing similar and relevant movie recommendations by analyzing the movie content, conducting multilabel genre classification based on movie overviews, and predicting movie ratings by utilizing various movie features.

For content based movie recommendation, two NLP approaches, namely Word2Vec and TF-IDF, are employed for capturing the similarities between the movie content and providing relevant movie suggestions based on similarity. Moreover, in multilabel genre classification tasks, various classification models are trained on the processed movie content and their performance of these models are evaluated and compared using precision, recall and f1-score as evaluation metrics to determine the effectiveness in classifying multiple genres for each

movie. Furthermore, various regression models are trained on preprocessed dataset to predict movie ratings and the model performances are compared utilizing mean squared error(MSE) and mean absolute error(MAE) as evaluation metrics.

The contributions and significance of the project are as follows:

- **Enhanced movie discovery:** Incorporation of NLP techniques for analyzing the movie content in movie recommendation system enhances the relevance of the suggestion and enriches movie discovery with user preferences.

- **Content Categorization:** Analyzing the movie description utilizing NLP techniques to identify the textual features indicative of specific genres is an useful approach for movie content categorization and this enables the effective classification of movies into multiple genres simultaneously.

- **Audience Insights:** Movie ratings prediction, along with exploratory data analysis, allows to identify the crucial factors to rating and viewer acceptance. Moreover, this analysis helps to understand the trends evident in audience preferences, which facilitates making strategic decisions to produce more engaging and successful movies.

References:

[1] Asaniczka. (2023), TMDB Movies Dataset, Retrieved from: https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies.