MOVIE PREDICTION

By:
PUNEETHRAJ K (4VP21MC030)
RASHMI V (4VP21MC033)

Chapter 1
INTRODUCTION

1.1 Introduction:

Movie is something emotional thing that may or may not affect to the human life. So, when we want to watch movies, it should be satisfying the most. It is necessary to give the ratings about the particular movie for the viewers. The ratings are the expressions about how people react about the movies. The other users can view these ratings about the movies. It will help the people to watch the movies. Also, it is helpful to the movie makers.

This movie prediction is tracker which is used to predict the movies for the users based on their age.

Movie reviews are an important way to gauge the performance of a movie. While providing a numerical/stars rating to a movie tells us about the success or failure of a movie quantitatively, a collection of movie reviews is what gives us a deeper qualitative insight on different aspects of the movie. A textual movie review tells us about the strong and weak points of the movie and deeper analysis of a movie review can tell us if the movie in general meets the expectations of the reviewer. This prediction algorithm will suggest the movie for the users based on their age. But we cal what the user's taste and what type of movie they watch. This prediction tells the accuracy of suggestion for the users and also movies.

1.2 Scope:

Many people watch movie to their own reason. It is one of the best way to find relaxation. So rating on movie is also important point to find best movie. Nowadays many platforms are available to book movies and also they provide a service to rate the movie. So by that viewer can get that particular movie rating. By using this preference dataset which provide the best suited movies for the people according to their age.

This movie prediction algorithm will help the people to watch the movies according to their age and the ratings that provide the best movies for watch.

1.3 Existing work:

Rating of movies are most important to movie makers and viewer. It helps to viewers to get to know about particular movie. Most of application are provide an option to rate the movie. In that can get rating of particular movie like 5 star, 4 star etc.

1.4 Proposed:

We all watch movies for entertainment, some of users never rate it, while some viewers always rate every movie they watch. This type of viewer helps in rating movies for people who go through the movie reviews before watching any movie to make sure they are about to watch a good movie. Here the viewers can get the reviews of most of the movies in at a time.

Here the we use the data about ratings and age of the user. These data are used for provide the best suited movie for the users based on the preference given by the viewers.

Row data is collected and using Logistic regression algorithm, it will provide the movies for the users.

Using this information, we can get all the ratings, preference and age of the particular users and movie details at one place.

1.5 Dataset:

We have already collect the dataset of type .csv from kaggle. [12]
This dataset contains the information about movie and user. For this we need to two dataset

1. One dataset contains the data about the movie ID, title, and genre of the movie
2. Second Dataset contains the user ID, movie ID, rating given by the user.
3. Third datset contains the data that contains all the above details as well as preference that is given by the user.

1.6 Objectives:

- Movie prediction used to predict the movies for the users based on the data provided by the users.
- Users can easily find the movies based on their age.
- Viewers can get the idea of how the movie is.
- Viewers can compare the ratings of one another.
- The users can easily find the movies of their taste based on the preference given by the viewed users of movie. It will match their age and provide the results.

Chapter 2
PROBLEM STATEMENT

Rating the movie after watching is most important now to other viewers to know about the movie as well as to movie makers to get to know about viewer's opinion about movie. They can predict the viewers interest.

This prediction will provide the movie genres based on their age. It will calculate the preference, ratings and userid of the particular user and provide the accurate result. But it cannot provide cent percent result. Because all the users does not have same mentality and taste. Based on the overall report it will provide the results.

This type of viewer helps in rating movies for people who go through the movie reviews before watching any movie to make sure they are about to watch a good movie. Analysing the rating given by viewers of a movie helps many people decide whether or not to watch that movie.

For this we need two files:

One dataset contains the data about the movie and the other dataset contains user data, who rated the movie. By combining these files, we get details about movie which is more rated. We can represent this data by using chart. That will give better visual. Also, we can list top rated movie.

The preference dataset that contains all the above details as well as preference that is given by the user. These preferences tell the viewers can watch this movie and it is preferred by the users who will watch it.

Chapter 3
LITERATURE SURVEY

The movie sector is one of the biggest contributors to the entertainment industry's unpredictability in success and failure. The aim of this research work to design an efficient movie recommendation algorithm that will increase prediction accuracy, the Movie Review Rating Prediction (MR2P) was achieved through a systematic review of the existing movie success algorithm. This research work will enable movie stakeholders (producers, directors, crew, cast already in the movie industry or aspirants) to know the kind of movie to invest in which will, in turn, be beneficial in terms of higher profit.

These are the reviews of some existing movie algorithms.

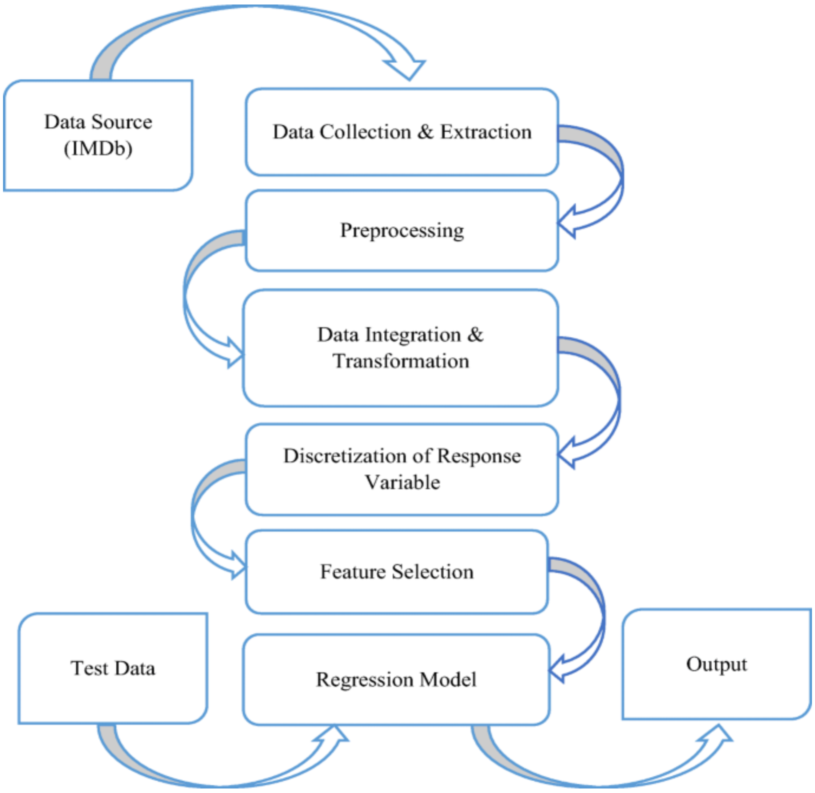
Linear Regression Algorithm (Shah et al., 2019)

The algorithm aimed to collect the number of likes, dislikes, and the view count of a trailer, release date, star ranking, and so on. Multiple Linear Regression Algorithm was used for the prediction of earnings of the movie

Algorithm (Steps of System Flow)

Input: Movie database, User input film with feature values

Output: Rating of user-entered film [1]


Data Source (IMDb)
Data Collection & Extraction
Preprocessing
Data Integration & Transformation
Discretization of Response Variable
Feature Selection
Test Data
Regression Model
Output

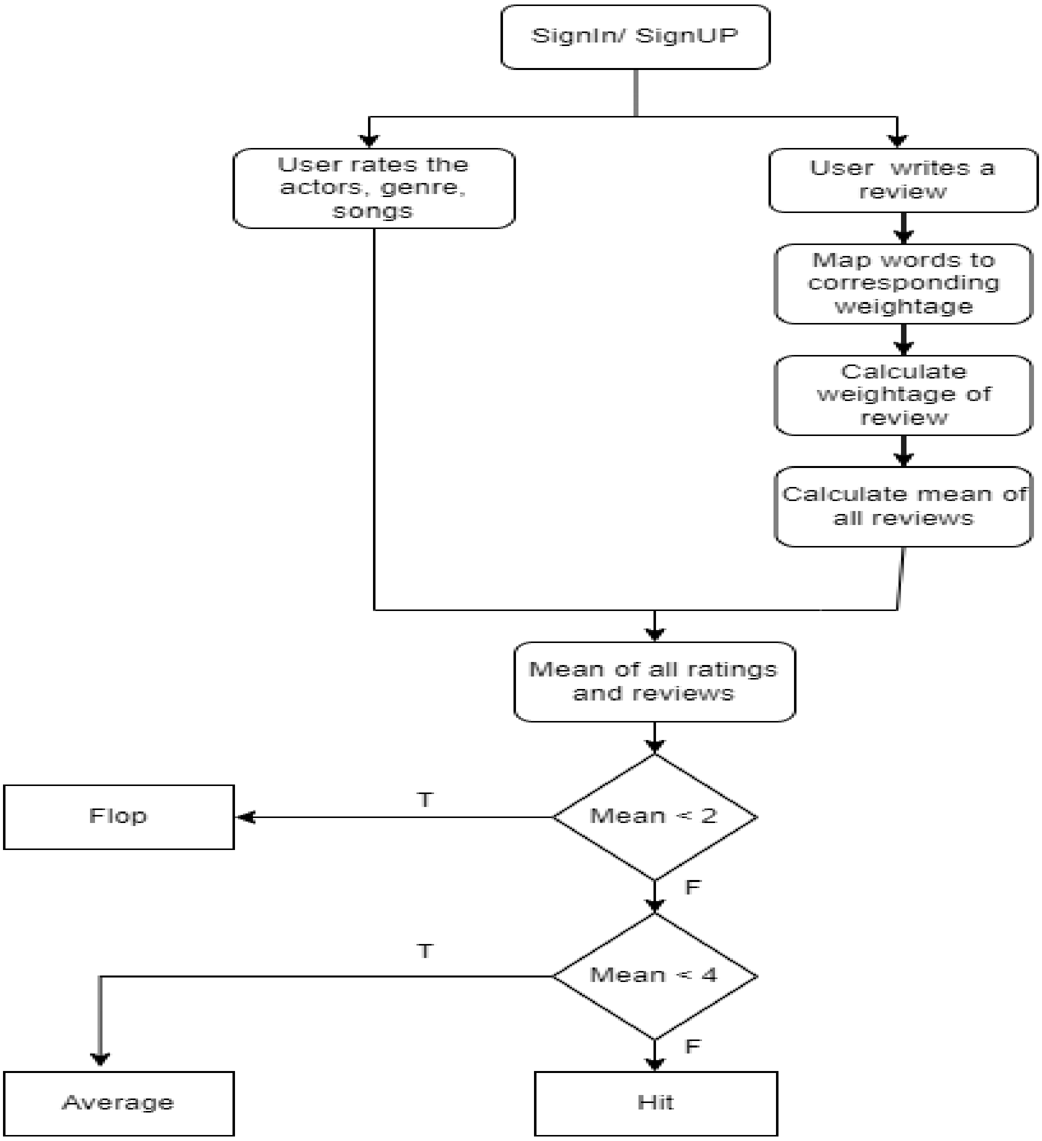
Weightage Algorithm (Antara, Nivedita, Shalin, Tanisha, & Pranali, 2018)

In this research, a custom website and algorithms for predicting the success class of a movie such as a flop, hit, or average was developed. In doing this, a custom dictionary of words in which common and important words used in reviews were stored according to the weightage assigned to them by the administrator. With the help of sentiment analysis, weightages will be assigned on a scale of one (1) to five (5), where 1 indicates the negative extreme, and five (5) indicates the positive extreme.

Algorithm (Steps of System Flow)

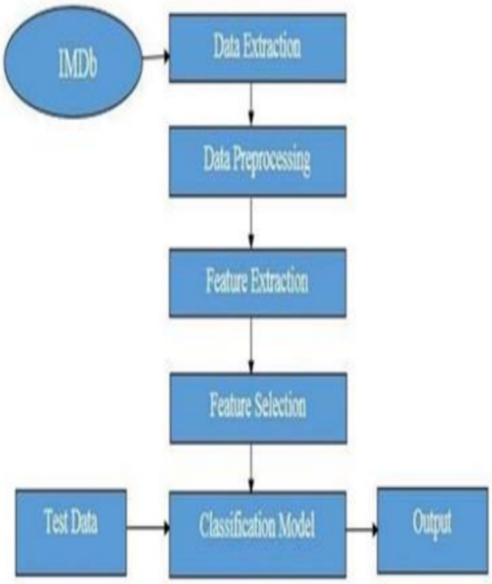
Input: User reviews and ratings for Actors, Genres and Songs

Output: Rating values in form of Emoji. [2]


SignIn/ SignUP
User rates the actors, genre, songs
User writes a review
Map words to corresponding weightage
Calculate weightage of review
Calculate mean of all reviews
Mean of all ratings and reviews
Flop
T
Mean < 2
F
T
Mean < 4
F
Average
Hit

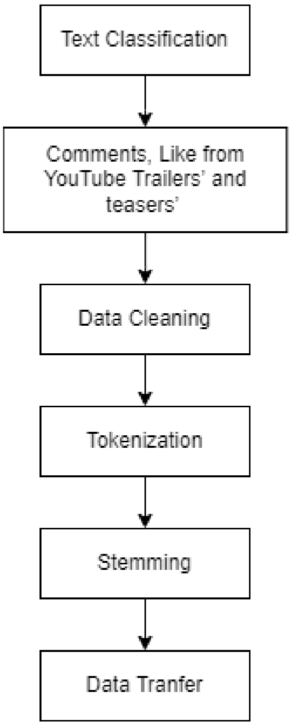
Random Forest by (Dhir & Raj, 2018)

The researchers wanted to dig deeper into the business side of movies and explore the economics behind what makes a successful movie. They wanted to examine the trends among films that lead them to become successful at the box office. [3]



Naïve Bays Algorithm

Naive Bayer's uses an identical method to predict the probability of various classes supporting various attributes. This algorithm is usually utilized in text classification. We used this Algorithm to predict the expected rating of a movie by using YouTube Trailers' and teasers' comments. [4]



In recent years, social media has become ubiquitous and important for social networking and content sharing. we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. [5]

Logistic Regression Algorithm

Logistic regression is a common algorithm used for binary classification problems, where the goal is to predict whether an observation belongs to one of two classes. In the context of movie prediction, logistic regression can be used to predict whether a movie will be successful (i.e., popular and profitable) or not.

To use logistic regression for movie prediction, we first need a dataset of movies. The target variable should be a binary variable

Once we have our dataset, we can train a logistic regression model using a training set of movies. The model will learn to identify patterns in the data that are associated with successful movies, and use those patterns to make predictions on new movies. [6]

We predict IMDb movie ratings and consider two sets of features: surface and textual features. For the latter, we assume that no social media signal is isolated and use data from multiple channels that are linked to a particular movie, such as tweets from Twitter and comments from YouTube. [7]

We consider the problem of predicting a movie's opening weekend revenue. Previous work on this problem has used metadata about a movie—e.g., its genre, MPAA rating, and cast—with very limited work making use of text about the movie. [8]

On-line news agents provide commenting facilities for readers to express their views with regard to news stories. The number of users supplied comments on a news article may be indicative of its importance or impact. [9]

Use of socially generated "big data" to access information about collective states of the minds in human societies has become a new paradigm in the emerging field of computational social science. [10]

Film Industry is not only a industry or a centre of entertainment, rather it is now a centre of global business. All over the world is now excited about a movie's box office success, popularity etc. We have used various movie list from Wikipedia and their rating from IMDb movie rating website to create the data set. [11]

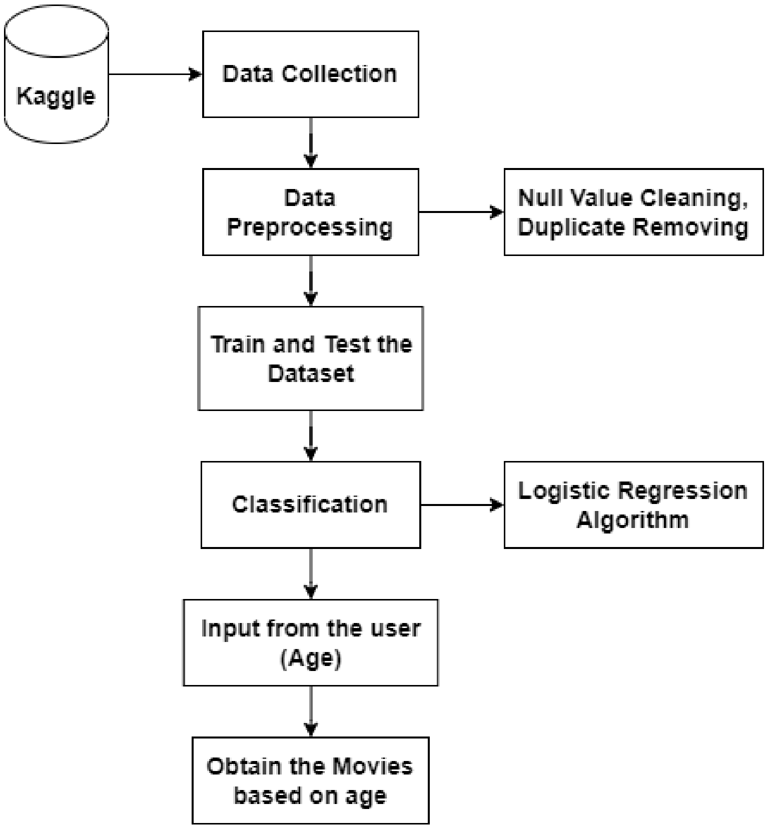
Movie studios often have to choose among thousands of scripts to decide which ones to turn into movies. Despite the huge amount of money at stake, this process—known as *green-lighting* in the movie industry—is largely a guesswork based on experts' experience and intuitions. [12]

Chapter – 4

ANALYSIS AND DESIGN

In this project, we collect the data from Kaggle website. It contains user information like userid and movie information like movie name, movie id and rating for the movie.

The preference dataset contains the information about the users, movies and the preferences that given by the viewers for the further watch.


Kaggle
Data Collection
Data Preprocessing
Null Value Cleaning, Duplicate Removing
Train and Test the Dataset
Classification
Logistic Regression Algorithm
Input from the user (Age)
Obtain the Movies based on age

Step 1: Data Collection

Collect data from Kaggle. Which contains movie information like movie name, release of year, movie id, genre and user information like userid, movieid.

Step 2: Data Pre – processing

Algorithm 4.1:

Step 1: In the pre – processing step, we are trying to cleaning the Null values.

Step 2: After that we are removing the duplicate values from Dataset.

Step 3: Classification

Algorithm 4.2: Logistic Regression Algorithm

Step 1: Data Pre-processing: This involves cleaning and transforming the raw data to prepare it for analysis. This includes handling missing values, dealing with outliers, and encoding categorical variables.

Step 2: Splitting the Data: The data is split into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate the model's performance.

Step 3. Feature Selection: Identify and select the independent variables that are most relevant to the outcome variable. This can be done using statistical tests, domain knowledge, or machine learning algorithms.

Step 4. Model Training: Fit a logistic regression model to the training data using the selected independent variables. This involves estimating the model parameters using maximum likelihood estimation.

Step 5: Model Evaluation: Evaluate the performance of the model using the testing set. Common evaluation metrics for logistic regression include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (ROC) curve.

Step 6: Model Improvement: If the model performance is not satisfactory, try improving it by adding more features, transforming the existing features, or changing the hyper parameters of the model.

Step 7: Model Deployment: Once the model is satisfactory, it can be deployed for making predictions on new data. It is important to monitor the model's performance over time and retrain it periodically to ensure that it remains accurate and up-to-date.

Chapter – 5
IMPLEMENTATION

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
import matplotlib.pyplot as plt
import warnings
from warnings import filterwarnings
filterwarnings("ignore")
movies = pd.read_csv('movies.dat', sep='::', engine='python', header=None, names=['movie_id', 'title', 'genres'], encoding='latin-1')
ratings = pd.read_csv('ratings.dat', sep='::', engine='python', header=None, names=['user_id', 'movie_id', 'rating', 'timestamp'])
users = pd.read_csv('users_old.csv', engine='python', header=None, names=['user_id', 'gender', 'age', 'occupation', 'zip_code'])
pref = pd.read_csv('preference.csv')
```

	user_id	movie_id	rating	timestamp	gender	age	occupation	zip_code	title	genres	preference
0	1	114508	8	1381006850	F	10	10	1.395649e+08	Species (1995)	Action|Horror|Sci-Fi|Thriller	1
1	1400	114508	7	1394296627	M	25	7	1.562806e+09	Species (1995)	Action|Horror|Sci-Fi|Thriller	0
2	1431	114508	8	1371699243	M	56	6	3.255869e+07	Species (1995)	Action|Horror|Sci-Fi|Thriller	0
3	2	499549	9	1376753198	M	56	16	1.752819e+07	Avatar (2009)	Action|Adventure|Fantasy|Sci-Fi	1
4	424	499549	8	1368291562	M	25	17	7.402026e+07	Avatar (2009)	Action|Adventure|Fantasy|Sci-Fi	1
...	...	...	...	...	...	...	...	...	...	...	...
101184	7992	1735190	7	1364009858	M	35	1	1.504685e+07	Ghett'a Life	Action|Drama	1

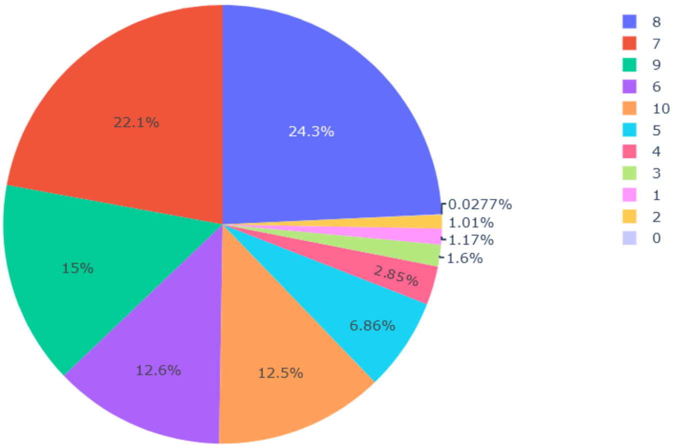
	user_id	movie_id	rating	timestamp	gender	age	occupation	zip_code	title	genres	preference
									(2011)		
101185	7998	2295564	3	1384731611	F	25	1	1.662540e+09	McCanick (2013)	Crime|Drama|Mystery|Thriller	1
101186	8000	489037	7	1390922879	M	45	7	3.576888e+08	Who Killed the Electric Car? (2006)	Documentary	1
101187	8000	1413496	8	1390941162	M	45	7	3.576888e+08	Revenge of the Electric Car (2011)	Documentary	1
101188	8000	2654360	8	1384844451	M	45	7	3.576888e+08	Deceptive Practice: The Mysteries and Mentors ...	Documentary	1

```
#Data Cleaning
df_dropped = pref.dropna()
```

```
#Count how many 10 ratings for the particular movie
topRate = pref.query("rating == 10")
topRate["title"].value_counts()
```

```
#Count of each rating
rating = pref["rating"].value_counts()
```

```
#Plot
rating = pref["rating"].value_counts()
numbers = rating.index
quantity = rating.values
import plotly.express as px
fig = px.pie(data, values=quantity, names=numbers)
fig.show()
```



```
#Training and testing the dataset
X = pref[['age','rating','user_id']]
y = pref['preference']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
y_pred

#Accuracy
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy:', accuracy)
Accuracy: 0.7501729419903153

#Based on age finding the movie
# Get the list of recommended movies for a new user based on their age
def get_recommendations(age):
    for i in age:
        user_data = [[i, 0, 0]]
        user_preference = model.predict(user_data)[0]# Set the user's ratings and ID to 0
        if user_preference == 1:
#           if pref[pref[age]] in :

            recommended_movies = pref[pref['age'] == i]['title']
            return list(recommended_movies)
        else:
            return []

# Example usage
new_user_age = int(input("Enter your age:"))
```

```
if new_user_age in range(1,8):
  agelist=list(range(1,8))
elif new_user_age in range(8,18):
  agelist=list(range(8,18))
elif new_user_age in range(18,25):
  agelist=list(range(18,25))
elif new_user_age in range(25,35):
  agelist=list(range(25,35))
elif new_user_age in range(35,50):
  agelist=list(range(35,50))
else:
  agelist=list(range(50,100))
recommended_movies = get_recommendations(agelist)
if recommended_movies:
    recommended_movies=list(set(recommended_movies))
    print('Top 10 Recommended movies for user with age', new_user_age, ':\n')
    for i in recommended_movies[0:10]:
        print(i)
else:
    print('No movie recommendation for the user who have the age',new_user_age,'.')
```

Chapter – 6

METHODOLOGY

We are using the Movies dataset for our project. In this we can get the data about the movies, users and preferences that is given by the movie viewer. We are implementing this project in Jupyter notebook for getting the result.

The movies dataset is .csv dataset, so that we can use pandas for reading this .csv dataset. Pandas is a Python library used for working with data sets. It has functions for analysing, cleaning, exploring, and manipulating data.

6.1 Data Pre - Processing:

Information pre - processing is an essential step for the creation of a machine learning model. To begin with, information may not be smooth or inside the required layout for the model which can reason misleading results. In pre - processing of information, we remodel information into our required format. It's for used to remove duplicates and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling etc.

In our project “Movie prediction” the pre - processing is done as follows:

1) Cleaned the missing values.
2) Removed Duplicated values.
3) Split the data into training and testing.
4) Building Logistic Regression Model.
5) Calculating accuracy score for model.

6.2 Classification

Logistic Regression is a classification algorithm that falls under the category of supervised learning. Its primary purpose is to predict the possibility of a target variable. In simpler terms, our dependent variable can only take two values, which are 1 (representing success/yes) or 0 (representing failure/no).

A logistic regression model is a mathematical algorithm that predicts the probability of a binary outcome based on input variables. This model is commonly used in machine learning to solve classification problems such as detecting spam, predicting the movies for the users etc. It is a straightforward and efficient method that can provide accurate results.

Logistic Regression has some rules as following:

- In case of binary logistic regression, the target variables must be binary always
- There should not be any multi-collinearity in the model, which means the independent variables must be independent of each other.
- We must include meaningful variables in our model.
- We should choose a large sample size for logistic regression.

To predict the “Movie prediction” we have used “Logistic Regression Algorithm” where we used movies dataset which have 101189 and 12 attributes. As the Logistic Regression rule says variable must be binary always. Then we implemented the Logistic Regression Algorithm.

6.3 Result

In our project “Movie Prediction” one algorithm was explored to predict the movie. This project hence helps to predict the Movie based on their age using the prediction model. Logistic regression algorithm gives the best accuracy compared to other Machine Learning Algorithms. Logistic Regression has an accuracy of 75.01%.

6.3.1 Accuracy

Algorithm	Accuracy
Logistic Regression	75.01%

6.4 Future Work

“Movie Prediction” is platform that suggest the best movies for the people based on their age. While age is an important factor in predicting movie preferences, other demographic factors such as gender, ethnicity, and income can also play a significant role. Future research could explore how incorporating these other factors can improve the accuracy of movie predictions.

As people age, their movie preferences may change. Future research could examine how preferences for different genres and types of movies change over time and how these changes can be incorporated into predictive models.

Social media can provide a wealth of information about people's movie preferences, such as their likes, shares, and comments on movie-related content. Future research could explore how incorporating social media data into predictive models can improve their accuracy.

Movie preferences can also be influenced by cultural factors such as language, religion, and nationality. Future research could examine how these factors can be incorporated into predictive models to improve their accuracy for specific cultural groups.

Conclusion:

In conclusion, a data analytics project focused on predicting movie preferences based on age using a movie dataset with preference data has the potential to provide valuable insights for the entertainment industry. By analysing the preferences of different age groups, we can identify patterns and trends that can inform marketing and content creation decisions.

Through the process of exploring the dataset, creating age groups, feature engineering, building predictive models, evaluating model performance, incorporating user feedback, and monitoring and updating models, we can build accurate predictive models that recommend movies to users based on their age and other relevant factors. This can improve the movie-watching experience for users and help movie studios and streaming services deliver more targeted and personalized content.

However, it is important to note that individual preferences can vary greatly and that age is just one factor among many that can influence movie preferences. Therefore, it is important to use a diverse and representative dataset, and to consider other factors such as gender, cultural background, and geographic location to get a more complete picture of movie preferences.

Overall, a data analytics project focused on predicting movie preferences based on age using a movie dataset with preference data can be a useful tool for the entertainment industry, but it should be used in combination with other data sources and with careful consideration of individual differences in preferences.

We use Kaggle dataset to get data

Kaggle is an online community platform for data scientists and machine learning enthusiasts.

Kaggle allows users to collaborate with other users, find and publish datasets.

References:

[1] Shah, K., Kapadia, J., Samel, Y., Saple, S., & Deshmane, P.. Movie Success Prediction using Data Mining and social media. International Research Journal of Engineering and Technology, 6(3), 188–190, (2019).

[2] Antara, U., Nivedita, K., Shalin, S., Tanisha, M., & Pranali, W.. Movie Success Prediction Using Data Mining. International Journal of Engineering Development and Research, 6(4), 198–203, (2018).

[3] Dhir, R., & Raj, A.. Movie Success Prediction using Machine Learning Algorithms and their Comparison. ICSCCC 2018 - 1st International Conference on Secure Cyber Computing and Communications, 385–390, (2018).

[4] Pirunthavi Sivakumar, Vithusia Puvaneswaren Rajeswaren, Kamalanathan Abishankar, E.M.U.W.J.B. Ekanayake, Yanusha Mehendran – Movie Success and Rating Prediction Using Data Mining Algorithms, 72-80, (2020)

[5] Asur, S., Huberman, B.A.: Predicting the future with social media. abs/1003.5699 (2010).

[6] Rachaell Nihalaani; Apoorva Shete; Darakshan Khan: 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (2021).

[7] Andrei Oghina, Mathias Breuss, Manos Tsagkias & Maarten de Rijke :Predicting IMDB Movie Ratings Using Social Media, Part of the Lecture Notes in Computer Science book series (LNISA,volume 7224) ISLA, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands.(2010)

[8]Joshi, M., Das, D., Gimpel, K., Smith, N.A.: Movie reviews and revenues: An experiment in text regression. In: Proceedings of NAACL-HLT (2010).

[9] Tsagkias, E., de Rijke, M., Weerkamp, W.: Predicting the volume of comments on online news stories. In: CIKM 2009, Hong Kong, pp. 1765–1768. ACM (2009).

[10] Márton Mestyán, Taha Yasseri , János Kertész: Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data, August 21, 2013.

[11] Warda Ruheen Bristi; Zakia Zaman; Nishat Sultana 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT): Predicting IMDb Rating of Movies by Machine Learning Techniques. (2019)

[12] Jehoshua Eliashberg, Sam K. Hui, Z. John Zhang: From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts, (2007).

[13] https://www.kaggle.com/datasets/tunguz/movietweetings?resource=download