MOVIE SUCCESS PREDICTION

A project dissertation submitted to Bharathidasan University
in partial fulfillment of the requirements
for the award of the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

Submitted by

TEJASWINI S
215214143

Under the guidance of
Mr. C.SATHISH KUMAR MCA.,M.PHIL.,SET,NET,
Associate Professor

PG DEPARTMENT OF COMPUTER SCIENCE (SHIFT-I)
BISHOP HEBER COLLEGE (AUTONOMOUS)
(Nationally Reaccredited at the 'A' Grade by NAAC with the CGPA of 3.58 out of 4)
(Recognized by UGC as "College with Potential for Excellence")
(Affiliated to Bharathidasan University)

TIRUCHIRAPPALLI 620017

APRIL 2023

DECLARATION

I hereby declare that the project work presented is originally done by me under the guidance of **Mr. C.Sathish Kumar,MCA.,MPhil.,NET,SET PG Department of Computer Science (Shift-I), Bishop Heber College (Autonomous), Tiruchirappalli-620 017**, and has not been included in any other thesis/project submitted for any other degree.

Name of the Candidate	**:**	**TEJASWINI S**
Register Number	**:**	**215214143**
Batch	**:**	**2021-2023**

Signature of the Candidate

PG DEPARTMENT OF COMPUTER SCIENCE (SHIFT-I)
BISHOP HEBER COLLEGE (AUTONOMOUS),
(Nationally Reaccredited at the 'A' Grade by NAAC with the CGPA of 3.58 out of 4)
(Recognized by UGC as "College with Potential for Excellence")
(Affiliated to Bharathidasan University)

TIRUCHIRAPPALLI - 620 017

Date:

Course Title: Project Course Code: P21CS4PJ

BONAFIDE CERTIFICATE

This is to certify that the project work titled **"MOVIE SUCCESS PREDICTION"** is a bonafide record of the project work done by **TEJASWINI S, 215214143,** in partial fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE** during the period **2021 - 2023.**

The Viva-Voce examination for the candidate **TEJASWINI S,215214143,** was held on _____.

Signature of the HOD **Signature of the Guide**

Examiners:

1.

2.

ACKNOWLEDGEMENT

I am grateful to GOD Almighty, who showered his blessings on me throughout this project and who has been an ineffable source of strength and inspiration in completing the project.

I am extremely thankful and indebted to **Dr. D. PAUL DHAYABARAN, M.Sc., M.Phil., PGDCA., Ph.D.,** Principal, Bishop Heber College (Autonomous), Tiruchirappalli, for providing me with the facilities and permission to carry out the project.

I pay my deep sense of gratitude to **Dr. K. RAJKUMAR, M.Sc., M.Phil., Ph.D.,** Associate Professor and Head, Department of Computer Science, Bishop Heber College (Autonomous), Tiruchirappalli, who encouraged me and provide me the opportunity to prepare the project.

I am grateful to **Mr.C.SATHISH KUMAR, MCA., M.Phil., SET, NET,** Associate Professor, Department of Computer Science (SFI), Bishop Heber College (Autonomous), Tiruchirappalli, the internal guide for timely suggestions and constant encouragement and support that led to the accomplishment of the project.

I would like to thank all my staff members of the PG Department of Computer Science (Shift-I). Who put their effort into shaping me. I am also immensely obliged to my friends for their elevating inspirational, support and encouragement in completion of this project.

Last but not the least, my parents, the one who aspires me to be focused and dedicated in the work I do, and also for being my moral support, so with due regards I express gratitude to them. Finally, I thank each and every one who has made a contribution towards the successful completion of my project.

TEJASWINI.S

ABSTRACT

Films are the most popular entertainment media in today's world. The total number of movies produced is growing at an exponential rate. This growth of movies utmost importance since billions of dollars are invested in making every movie. Before the movie is released the success of the movie relies on media hype, previews and pre-release marketing drive. But these don't determine the movie success when it is released. The main idea aims to explore more machine learning algorithms and use them to successfully predict movie earnings and ratings using kaggle movie dataset. Our goal is to rank movies based on their social media popularity, whether they are flop, hit or superhit. It is based on the investment of the film from the domestic box office earned through the film. Predictions are made immediately after the film premieres and are more accurate. IMDb (Internet Movie Database) contains categorical and numerical information such as IMDb score, Director, Producer, Actor, Gross and Budget. Before it hits theatres, it provides a way to predict the movie success by hearing from critics and others whether a film is successful or not. K- Nearest Neighbours Regressor algorithm are employed in this project to predict the financial market of various movies from inputs. The algorithm executes a perfect work of prediction and shows better accuracy. To improve the interpretability and performance of machine learning algorithms feature selection is widely used. There also have been various semantic analysis techniques to analyze user reviews which were applied to analyze the IMDb movie ratings. This system makes it easier for the users to book the tickets in advance.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Background and Motivation

Outside of the crew's popularity and movie popularity, the success of a movie is highly dependent on digital forces like online campaign, marketing, reviews, box office conditions and opening day ticket sales. Though, before a movie is released, the success of the movie has to rely on media hype, previews and prelease marketing drive, but these don't determine or translate to a movie's success when released. The problem faced is most of the producers, directors and stakeholders end up expending millions on movie budget without knowing if the movie would be a success or a failure. This study proposes a model that analyzes the revenues generated from previous movies, the reviews, ticket sales, crew's popularity and marketing budget to predict the success of a movie. Such a prediction could be veritably useful for motion studios to make better intelligent opinions like perfecting content and creativity, artist compensations, advertising and marketing of the movie consequently. This helps investors and stakeholders to predict an accurate return- on-investment(ROI) and other associated profit or loss.

1.2 Objectives and Scope of the Project

Movie success prediction will help producers understand that the investment made by them in the movie is going to be worth it or not. Data mining and machine learning algorithms will analyze all the past data of the components and by using various machine learning algorithms the model can predict the result of the movie. Past performance, financial plan, releasing date, actress, director and actor will contribute the predictions based on by training the model with the past data. Adding dynamic factors to be more precise about the prediction, twitter tweets and hash tags are analyzed on the basis of sentiments. This sentiment analysis returns positive, negative and neutral hype as results for the movie. The main aim of Movie Success Prediction using machine learning is to

propose a system that helps to predict the success of movies. This will predict whether the movie has been flop or hit or super hit based on random forest algorithm. Project emphases on devising and develops a mathematical model to predict the movie performance. The mathematical approach for prediction of successful movies is the major idea of the project.

CHAPTER 2

RELATED WORK

2.1 Comparison of Existing Systems

Partha Chakraborty, et.al.,**[1]** proposed the model using Linear Regression and Data mining Techniques.Predicting the success of the movie the society's reaction to a new product in the sense of popularity and adaption rate has become an emerging field of data analysis . There are a huge number of data about movies is available in the web. By studying these data anyone can find connection between some attributes of movies over their success. These factors can be used by producers and production houses to make their creation profitable. In this work, data mining technique will be used to extract patterns which can be useful in anticipating films success. The researchers used variety of machine learning models includes Linear Regression and data mining techniques based on historical data.

Narayana Darapaneni, et.al.,**[2]** proposed the system using Random Forest, DecisionTree, K-Nearest Neighbours (KNN), NLP, XGBoost Classifier and Deep Neural Network. The researchers used a variety of machine learning models, such as Random Forest, Decision Tree, K-Nearest Neighbours (KNN), NLP, XGBoost Classifier and Deep Neural Network were performed. They were implemented on IMDB dataset for predicting Success of movies. The solution to predict the success rate of a movie is by performing predictive analysis on the various features of the movie. The model predicts the Success, based on different attributes or features of the movie. Movie crew (including director producer, music director), Movie plot (Storyline), Box-Office revenue, Audience and Critics reviews / ratings. Based on the results, XGBoost Classifier gives the best accuracy.

S.Sahu, et.al., proposed **[3]** the model by using Collaborative filtering, Content-based filtering.This research paper proposed content-based movie recommendation system by using tmdb_5000_movies and tmdb_5000_credits datasets, which are publicly available. The movie hit prediction and target audience prediction module make use of the IMDb rating dataset. The

tmdb_5000_movie data set is consistent with 4803 movie data. The tmdb_5000_credits data set consisted of 4813 movie data. The data set has 4 attributes movie_id, title, cast, and crew. Content-based movie recommendation system model produces 10 most similar movies of all the movies listed in the data set. his study has used publicly available Internet Movie Database (IMDb) data and The Movie Database (TMDb) data. The multiclass classification model is implemented and achieved 96.8% accuracy, which outperforms all the benchmark models. This study highlights the potential of predictive and prescriptive data analytics in information systems to support industry decisions.

Olubukola D. Adekola, et.al., proposed **[4]** the model by using the algorithms used are: Naïve Bayes and Multi-layer Perception, K-fold cross-validation and K-Nearest neighbours (KNN).The movie success prediction model is a method for reading and learning the data contained in the IMBD dataset was implemented. The input data for each movie would be a vector of three dimensions containing the average score of its writers, actors and directors. The score for each actor, writer, director is a weighted sum of the scores of all the movies each of the dimensions have been involved in directly and indirectly. This included various techniques such as data preprocessing and feature reduction which also promoted accuracy. The main part of Data mining is to detect the patterns. The predictor calculates the output score of a movie from the input vector using weighted sum of k-nearest neighbours (KNN). The weights are optimized with cost function built by K-fold cross-validation. This study has been able to address movie success prediction while it could be extended to tackle other related issues in the movie industry.

Dewan Muhammad Qaseem, Nashit Ali, et.al., proposed **[5]** the model is focused on developing a technique to predict movie success rates based on viewers' tweets on movie trailers. It predicts the movie rating in the star's form (1-5). By applying the hash tag method tweets are collected from several movies once the trailer was released. We have used four key algorithms (Naïve Bayes, SVM, decision tree, and KNN) on NLTK Movie review corpora and train & test our models. Machine learning training data sets were not readily available for movies rating; then, we shifted towards a lexicon-based approach. All these three dictionaries have a different word

count, and each word in these dictionaries has its own polarity in the form of a score. Finally, we have also compared our results with other movie rating sites like IMDB rating, which are satisfactory.

Tanishq Sharma, et.al., **[6]** model uses the algorithms used are Random Forest Classifier. The research paper made an achievement pace of a film which is mainly dependent on the marketing strategies the producers use on social media. Predicting the success of a movie using Machine Learning concepts automatically serve their purpose. It helps lot of cinema hall owners as they need to manage screening for the movie. In this paper, we are fetching comments and tweets from YouTube and Twitter API after the release of the trailer of the movie and performing Sentiment analysis on them. Thus, we propose a system to predict box-office movie success using machine learning concepts by analyzing the sentiments of movie-related comments from Twitter and YouTube.

Hrithik Jain, et al. **[7]** proposed , the model using the data mining techniques**.** The main focus of our project is to predict whether the movie will be successful, i.e. hit or unsuccessful, i.e. flop, based on certain attributes (static and social/dynamic). Few of the static factors included for determining whether the movie will be successful or not are: genre, film budget, actors, actresses, director, producer, production house, release date etc. From the social media perspective, i.e. dynamic, we would be looking at trending hashtags on Twitter related to the movie. The dynamic components of twitter analysis provide prediction of social media success or failure value. Simple UI and non technical user experience have been achieved for both types of users i.e creator and watcher of the movie. Accuracy of 85% has been observed in the project for predictions of movies yet to be released on the basis of algorithms and twitter sentiment analysis.

Dipak Gaikar ,et.al., **[8]** developed a mathematical model for predicting the rating and success classes such as hit, flop and neutral of the movies. In order to do this we have used a machine learning and data mining algorithm. The algorithm used for classification is k-NN. Popularity factor of various movie parameters like actor, actress, director, writer, budget etc. is collected which helps in the movie success prediction. This project helps the director or producer

of the movie to pre-decide the parameters such as actor, actress etc. of the movie. The original data is gathered by referring to the number of followers of actors, actresses, directors, and writers on social media sites like Facebook, Twitter, Instagram. In this research paper, data are gathered from the IMDB ratings of past movies from 1998 to 2018. This project follows a mathematical model of k-N N algorithm which is used for classification. It searches for "k" nearest neighbors . It performs classification by classifying a case by a majority vote of its neighbours. Users can easily decide to view which content and stake holders can confidently finance into the correct movie project.

M. Ahmed, et.al., proposed **[9]** to predict the movies popularity" describes experiments in predictive analysis using machine learning algorithms on both conventional features, collected from movies databases on Web as well as social media features (text comments on YouTube, Tweets). The results demonstrate that the sentiments harnessed from social media and other social media features can predict the success with more accuracy than that of using conventional features. We achieved best value of 77% and 61% using selected social media features for Rating and Income prediction respectively, whereas selected conventional features gave results of 76.2% and 52% respectively.

Saurabh Kumar, et.al., proposed **[10]** paper contains data mining technique and machine learning algorithms using R software to predict the success and failure of movie based on several attributes. The dataset requirement for this paper is fulfilled through kaggle repository. This dataset consists of 651 rows and 32 columns. The prediction of 33.94 audience score with 95% confidence level that the score between 5.62 and 62.26. The actual audience score of this movie is 19. The predicted audience score is greater than actual audience score. In this research paper, movies success is predicted on the basis of critics score. Several attributes can be included as predictors and build model for that attributes to perform prediction. If there was a movie gross score and movie net profit along with movie manufacturing cost, then build a more strong model for movie success prediction.

2.2 Overview of the Proposed System

One of the most frequently considered approaches to improving the interpretability and performance of machine learning algorithms is feature selection. Our proposed system is based on the fact that the features used in our study have already been shown to be very good at predicting a movie's success in previous studies. Also keep in mind that the features we choose should fit the process of building the predicting model. Additionally, we have taken into account the fact that the study's use of fewer features has a negative impact on a prediction model's performance. We will entirely rely on machine learning algorithm for prediction, which is used as the K-Nearest Neighbours Regressor predicted results will be validated .Finally, the model and involving it for the significant perspective that is for expectation of film industry outcome of another film before its delivery. There are currently a number of approaches designed to determine a movie's prediction. For a business decision support system, some of our methodologies incorporate statistical analytics and machine learning tools.

CHAPTER 3

DATA COLLECTION

3.1. Source of Data

Data collection is the most efficient method in machine learning. It involves gathering relevant data from various sources, cleaning, preprocessing and organizing in to a format suitable for analysis. Data is collected and measured from different resources in IMDB 5000 Movie Dataset (Internet Movie Database) dataset.

Source: https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset

IMDb (Internet Movie Database) is a popular online database of information related to films, television programs, and video games. It contains a vast collection of data on movies, including information on cast and crew, plot summaries, user ratings, and box office performance. The dataset has 5042 rows and 20 columns. The data is extracted from IMDB site with movies released in the years between 1916 and 2016.
All column headings are identical and ratings are collected from different people reviews. The scraper is coded in basic extraction system and sequentially downloads the information for each movie in Spyder environment.

3.2. Description and Exploration of Data

Primary data taken from the online sources which remains in the form of raw statements, digits and qualitative terms. There are faults, exclusions and inconsistencies in the impure data requires corrections. To operate machine learning requires two things, one is data and another is models. The dataset contains 5042 rows and 20 columns. There are totally 4533 unique name of the director, and thousands of actors . The variable “imdb_score” acts as the response variable, the other 19 variables acts as the possible predictors. The dataset is collected from IMDb website which is employed through training the model. Availability of data is in various formats. This can

be need to convert it into numerical format in order to apply machine learning algorithm effectively. It is eminent as it comprises of movies information and their commentaries from movie viewers. IMDb scores are greatly standard and suggested by public. This dataset includes information such as Year, Movie Title, Director Name, Actor Name, Total Facebook Likes, Genres, Number Of User Reviews, Number Of Critic Reviews, Duration, Gross, Number Of Voted Users, Language, Country, Content Rating, Budget, IMDB Score, Aspect Ratio, Face number in Poster, Plot Keywords, IMDB Movie Link.

Categorical Columns:

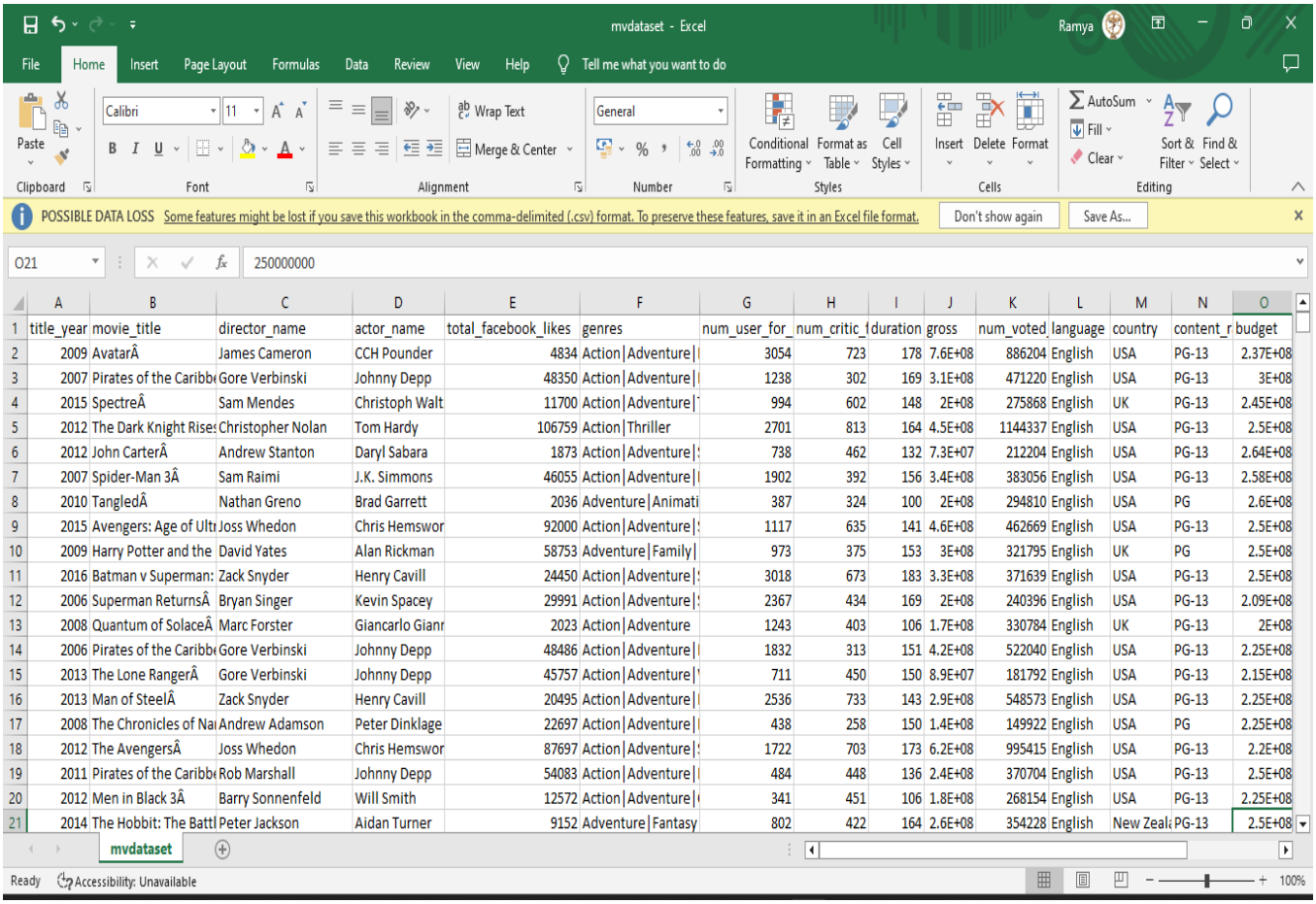
movie_title, director_name, actor_name, genres, language, country, content_rating, plot_keywords, movie_imdb_link

Numerical Columns:

title_year, total_facebook_likes, num_critic_for_reviews, num_user_for_reviews, duration, gross, num_voted_users, budget, imdb_score, aspect_ratio, facenumber_in _poster

FEATURES	DESCRIPTION
title_year	The year in which the movie is released.
movie_title	Title of the movie.
director_name	Name of the director of the movie.
actor_name	Name of the actor of the movie.
total_facebook_likes	Total number of facebook likes of the entire cast of the movie
Genres	Film Categorization like Adventure, Horror, Sci-Fi, Romance, Comedy, Action, Family

num_user_for_reviews	Number of users who gave reviews.
num_critic_for_reviews	Number of critical reviews on IMDB.
Duration	Duration in minutes.
Gross	Gross earnings of the movies in rupees.
num_voted_users	Number of people who voted for the movie.
Language	English, Arabic, Chinese, Japanese, French.
Country	Country where the movie is produced
content_rating	Content rating of the movie
Budget	Budget of the movies in rupees
imdb_score	IMDB score of the movie on IMDB.
aspect_ratio	Aspect ratio the movie was made in.
facenumber_in_poster	Number of the actor who featured in the movie poster
plot_keywords	Keywords describing the movie plot.
movie_imdb_link	IMDB link of the movie.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	title_year	movie_title	director_name	actor_name	total_facebook_likes	genres	num_user_for_	num_critic_f	duration	gross	num_voted_	language	country	content_r	budget
2	2009	AvatarÂ	James Cameron	CCH Pounder	4834	Action|Adventure|	3054	723	178	7.6E+08	886204	English	USA	PG-13	2.37E+08
3	2007	Pirates of the Caribb	Gore Verbinski	Johnny Depp	48350	Action|Adventure|	1238	302	169	3.1E+08	471220	English	USA	PG-13	3E+08
4	2015	SpectreÂ	Sam Mendes	Christoph Walt	11700	Action|Adventure|	994	602	148	2E+08	275868	English	UK	PG-13	2.45E+08
5	2012	The Dark Knight Rises	Christopher Nolan	Tom Hardy	106759	Action|Thriller	2701	813	164	4.5E+08	1144337	English	USA	PG-13	2.5E+08
6	2012	John CarterÂ	Andrew Stanton	Daryl Sabara	1873	Action|Adventure|	738	462	132	7.3E+07	212204	English	USA	PG-13	2.64E+08
7	2007	Spider-Man 3Â	Sam Raimi	J.K. Simmons	46055	Action|Adventure|	1902	392	156	3.4E+08	383056	English	USA	PG-13	2.58E+08
8	2010	TangledÂ	Nathan Greno	Brad Garrett	2036	Adventure|Animati	387	324	100	2E+08	294810	English	USA	PG	2.6E+08
9	2015	Avengers: Age of Ultr	Joss Whedon	Chris Hemswor	92000	Action|Adventure|	1117	635	141	4.6E+08	462669	English	USA	PG-13	2.5E+08
10	2009	Harry Potter and the	David Yates	Alan Rickman	58753	Adventure|Family|	973	375	153	3E+08	321795	English	UK	PG	2.5E+08
11	2016	Batman v Superman:	Zack Snyder	Henry Cavill	24450	Action|Adventure|	3018	673	183	3.3E+08	371639	English	USA	PG-13	2.5E+08
12	2006	Superman ReturnsÂ	Bryan Singer	Kevin Spacey	29991	Action|Adventure|	2367	434	169	2E+08	240396	English	USA	PG-13	2.09E+08
13	2008	Quantum of SolaceÂ	Marc Forster	Giancarlo Gianr	2023	Action|Adventure	1243	403	106	1.7E+08	330784	English	UK	PG-13	2E+08
14	2006	Pirates of the Caribb	Gore Verbinski	Johnny Depp	48486	Action|Adventure|	1832	313	151	4.2E+08	522040	English	USA	PG-13	2.25E+08
15	2013	The Lone RangerÂ	Gore Verbinski	Johnny Depp	45757	Action|Adventure|	711	450	150	8.9E+07	181792	English	USA	PG-13	2.15E+08
16	2013	Man of SteelÂ	Zack Snyder	Henry Cavill	20495	Action|Adventure|	2536	733	143	2.9E+08	548573	English	USA	PG-13	2.25E+08
17	2008	The Chronicles of Nar	Andrew Adamson	Peter Dinklage	22697	Action|Adventure|	438	258	150	1.4E+08	149922	English	USA	PG	2.25E+08
18	2012	The AvengersÂ	Joss Whedon	Chris Hemswor	87697	Action|Adventure|	1722	703	173	6.2E+08	995415	English	USA	PG-13	2.2E+08
19	2011	Pirates of the Caribb	Rob Marshall	Johnny Depp	54083	Action|Adventure|	484	448	136	2.4E+08	370704	English	USA	PG-13	2.5E+08
20	2012	Men in Black 3Â	Barry Sonnenfeld	Will Smith	12572	Action|Adventure|	341	451	106	1.8E+08	268154	English	USA	PG-13	2.25E+08
21	2014	The Hobbit: The Battl	Peter Jackson	Aidan Turner	9152	Adventure|Fantasy	802	422	164	2.6E+08	354228	English	New Zeal	PG-13	2.5E+08

Fig. 3.1 Dataset Movie Success Prediction Using Machine Learning

CHAPTER 4

PREPROCESSING AND FEATURE SELECTION

4.1 Preprocessing Methods and Steps

Data Preprocessing is a process converts the raw data into a clean data set. It helps in transformation process of raw data into an clear format. It is an important step in our project. Data quality needed to be checked before applying machine learning algorithms.

- **Data cleaning :** Removing the incorrect data, partial data and imprecise data from datasets, and helps in replacing the missing values.
- **Imputing Mean values**-Null values are replaced as the mean values in the dataset..
- **Data transformation :** Changes made in the data format or on the structure of the data is data transformation. It can be simple or complex based on the requirements.
- **Data reduction :** Reduces the data volume to make the analysis easier. It helps to minimize the storage space. Dimensionality reduction technique is implemented.

In our project we are implementing the scaling preprocessing Standard Scaler and Min Max Scaler with data transformation preprocessing method. The null values are removed by using the drop method and we are checking the null values. Few variables are in the range of millions and in tens, to bring them in to same scale data transformation preprocessing method is implemented.

4.2. Feature Selection

Concluding the subset of features is called feature selection. A feature vector is calculated based on the algorithm's input data which is too huge to be processed or it is redundant. Dimensionality reduction is achieved by feature extraction. Instead of the complete initial facilitates of data to carry out the desired operation. To find the optimal features machine learning model training can sometimes be a ridiculous task to accomplish. The input features are taken from the IMDB 5000 Movie Dataset such Content rating, Year, Movie name, Director name, Actor name, Genres, Gross and Total facebook likes.

CHAPTER 5

METHODOLOGY AND SYSTEM DEVELOPMENT

5.1. Proposed Methodology

The proposed system is to develop a model for prognosticating the success of movie being a Flop or Hit using machine learning techniques and algorithms. After collecting the relevant data, data must be cleaned and prepared for model development. The first step is to identify a dataset of film data which is required for analysis. Applicable attributes need to be named from the movie data. Attributes can be general pre-product information regarding film products analogous as movie title, effect, order, language and information about actors, and directors. Also, the data must include some measure of success, analogous as user movie conditions. The applicable dataset must be made and structured in such a way that the data used is representative of the movie scene at large, as well as suitable for analysis by the applicable machine learning ways and algorithms. To increase the accuracy and efficiency of this task, the data is divided in to training and testing data with a ratio of 80% training data and test data is of 20%. Further, correlation is performed on applicable dataset to find the relationship between al the variable with each other. The model prediction performance uses proposed machine learning algorithm which must be estimated on the dataset in order to determine success and failure of movie directly. Random Forest Regressor, Gaussian Process Regressor and K- Nearest Neighbours Regressor algorithm are the regression algorithms used in this project.

The proposed system shows out a process design of execution system

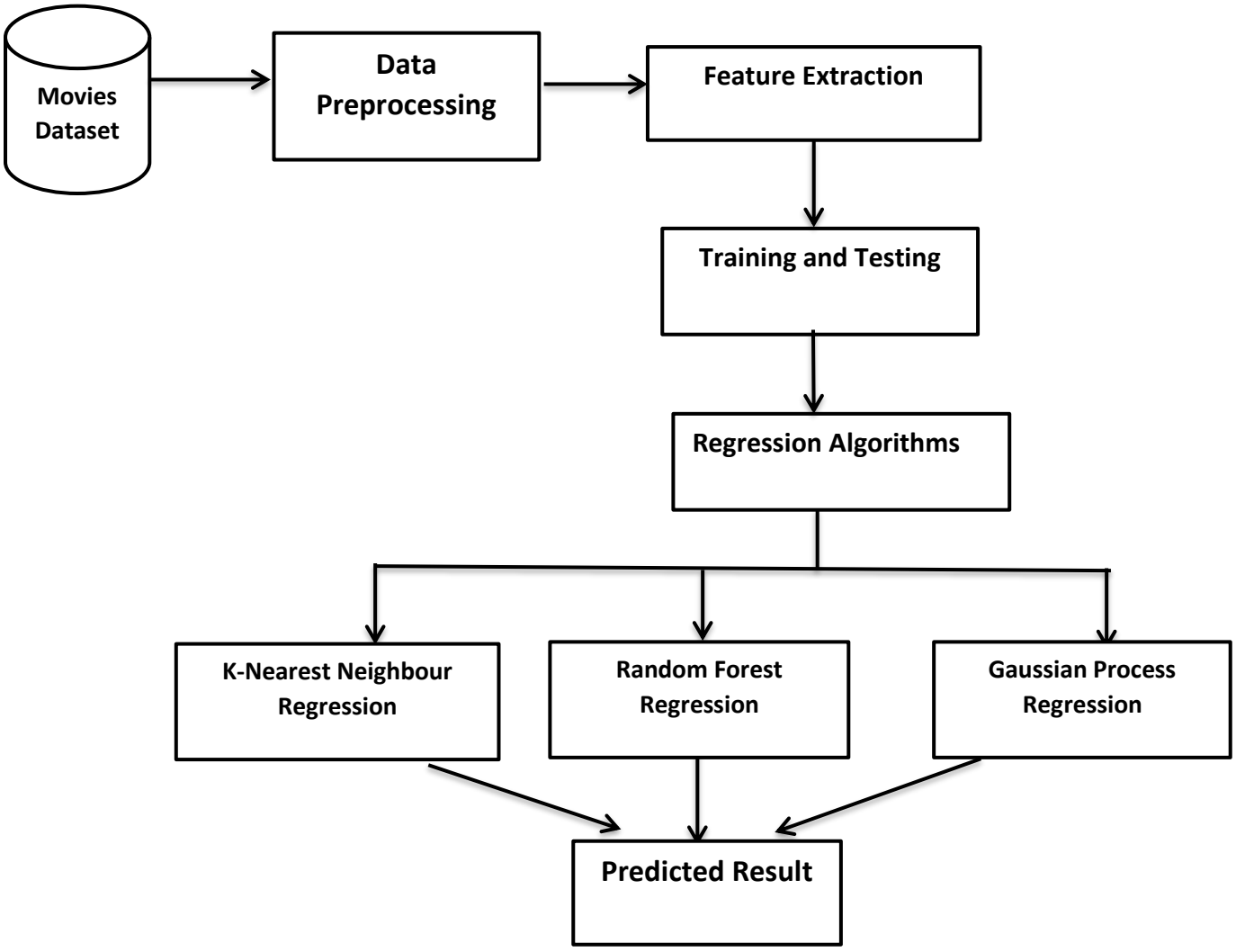


Fig. 5.1 Movie Success Prediction Using Machine Learning

5.2. Algorithms, Training and Testing

K-Nearest Neighbor Regression:

K Nearest Neighbors (KNN) Regression is a machine learning algorithm that is used for regression tasks. The KNN Regression algorithm works by finding the K training instances that are closest to the new instance to be predicted, based on a similarity metric such as Euclidean distance or cosine similarity. It can handle non-linear relationships between the features and the target variable. It can handle missing data and outliers in the dataset. The algorithm then takes the average or median of the target values of these K nearest neighbors as the predicted value for the new instance. The value of K is a hyperparameter that needs to be tuned based on the dataset and the problem at hand. The performance of KNN Regression may be sensitive to the choice of similarity metric and the scaling of the features.

Training And Testing:

The dataset is divided into a training set and a testing set. The IMDB 5000 movie dataset split into 80% of training and 20% of testing. First subset is referred as training data, then the actual dataset is used to train a machine learning model. The KNN Regression algorithm works by finding the K training instances that are closest to the new instance to be predicted. The input features are taken from the IMDB 5000 Movie Dataset such Content rating, Year, Movie name, Director name, Actor name, Genres, Gross and Total facebook likes. The model is selected and trained with the available data. The model is trained and validated; it can be used to predict the dependent variable for new data. This is done by inputting the values of the independent variables into the model and obtains the predicted value Hit or Flop based on the dependent variable(imdb_score).

Random Forest Regression:

Random Forest Regression is a supervised machine learning algorithm that is used for regression tasks. It can handle missing data and outliers in the dataset. It belongs to the family of ensemble methods, which means that it combines the predictions of multiple models to improve the overall accuracy and stability of the prediction. The Random Forest

algorithm is based on decision trees. The goal of the algorithm is to construct an ensemble of decision trees that work together to make a prediction. This ensemble method reduces the variance of the model and improves its generalization performance. During the training process, the algorithm selects a random subset of features at each node of each tree, and then chooses the best feature and the best split point among them. This process is repeated for each tree in the forest.

Gaussian Process Regression:

Gaussian Process Regression (GPR) is a probabilistic machine learning algorithm that models the distribution over functions, rather than a single function, to make predictions. The algorithm assumes that the target variable y is a realization of a Gaussian process, which is a collection of random variables such that any finite number of them have a joint Gaussian distribution. The GPR algorithm uses Bayes' rule to compute the posterior distribution over functions, given the observed data. The mean function of the Gaussian process is typically assumed to be zero, and the covariance function is usually specified by a kernel function k(x, x') that measures the similarity between two inputs x and x'. The most commonly used kernel function is the radial basis function (RBF) kernel, which is defined as:

$$\mathbf{k(x, x') = \sigma f^2 \exp(-(x - x')^2 / (2\ell^2))}$$

where σf is the amplitude parameter, ℓ is the length scale parameter, and $(x - x')^2$ is the squared Euclidean distance between the inputs x and x'. Handles non-linear relationships between the features and the target variable, and can capture complex patterns in the data.

CHAPTER 6

SYSTEM EVALUATION

6.1. Summary of Evaluation Metrics and Methods

The success of the movie was analyzed based on the five metrics:

- Mean absolute error (MAE),
- Mean squared error (MSE),
- Root mean squared error (RMSE),
- Accuracy (Acc).

Mean Absolute Error (MAE):

MAE stands for Mean Absolute Error, and it is a common metric used to evaluate the performance of regression models in machine learning. MAE is calculated as the average of the absolute differences between the predicted and actual values of a continuous variables. It is often used in combination with other metrics like RMSE (Root Mean Squared Error) and R-squared to evaluate the performance of a regression model.

MAE = mean(abs(predicted - actual))

Where:

predicted: the predicted value of the target variable

actual: the true value of the target variable

mean: the average value of the absolute differences between predicted and actual values.

Mean Squared Error(MSE):

MSE is calculated as the average of the squared differences between the predicted and actual values of a continuous variable. It provides a measure of how well the model fits the data by measuring the average squared distance between the predicted and actual values.

MSE = 1/n * sum((predicted - actual)^2)

Where:

predicted: the predicted value of the target variable

actual: the true value of the target variable

n: the number of data points in the dataset.

Root Mean Squared Error:

RMSE stands for Root Mean Squared Error, is a common metric used to evaluate the performance of regression models in machine learning. RMSE is calculated as the square root of the average squared difference between the predicted and actual values of a continuous variable.

RMSE = sqrt(mean((predicted - actual)^2))

Where:

predicted: the predicted value of the target variable

actual: the true value of the target variable

mean: the average value of the squared differences between predicted and actual values.

Accuracy:

Accuracy is a metric that describes how the model works across all classes in general, when all courses are equally essential. Accuracy is calculated by dividing the overall number of predictions by the number of correct predictions.

Accuracy = Total number of correct predictions / total number of elements

In these five metrics are performed well, to improve the model and to predict the movie whether it is hit or flop. However, K-Nearest Neighbour Regression, Random Forest Regression and Gaussian Process Regression are the machine learning models used.

6.2. Experimental Results and Discussion

In this work, experimental results suggest that movie success prediction models can compare the highest accuracy of three machine learning models. In this project the success of the movie is predicted to find whether the predicted movie is flop or hit based on the imdb_score. Splitting the training and testing data using machine learning algorithms are K-Nearest Neighbour Regressor, Random Forest Regression and Gaussian Process Regression. In this project dataset is divided into training (80%) and testing (20%). This result show that the K-Nearest Neighbour Regressor models performed the best on the hold-out test, with the evaluation metrics of MAE, MSE, RMSE, Accuracy.

ALGORITHMS	MAE	MSE	RMSE	ACCURACY
K-Nearest Neighbour Regressor	1.0784	1.832	1.353	0.981
Random Forest Regression	0.911	1.384	1.176	0.802
Gaussian Process Regressor	0.194	0.106	0.326	0.794

Fig. 6.2.1 Table: Comparison of Machine Learning Algorithms

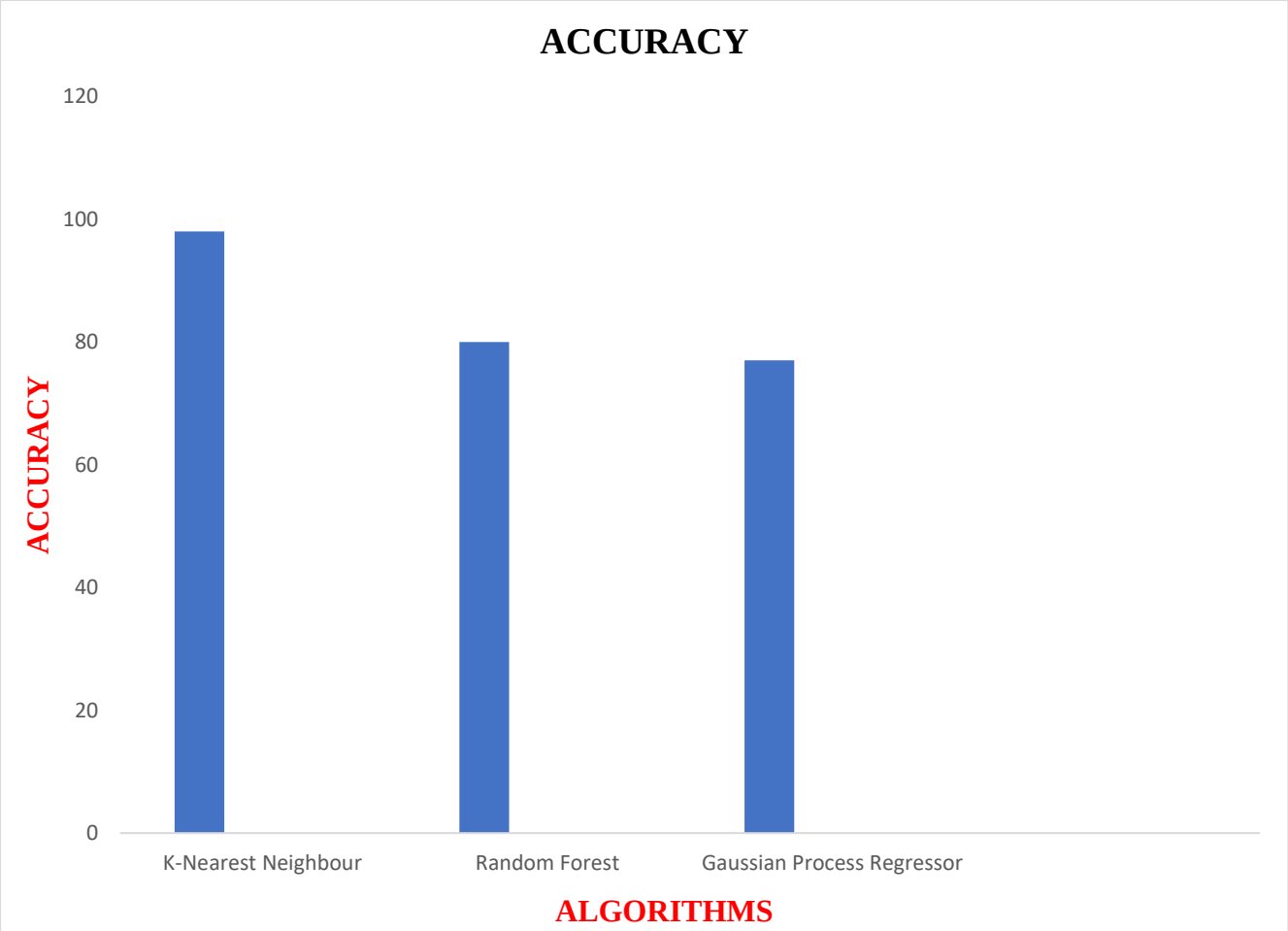


Fig. 6.2.2 Graph: Accuracy Comparison of Machine Learning Algorithms

The above diagram describes that are used in Movie success prediction and compares the results of the K-Nearest Neighbour Regressor, Random Forest Regression and Gaussian Process Regression. The K-Nearest Neighbour Regressor outperforms other ML regressions in terms of accuracy. Evaluation metrics such as mean absolute error, mean squared error, root mean squared error and accuracy were used to analyze the results. The accuracy of the model determines its performance. Predicting the success of the movie analyses the accuracy results of the K-Nearest Neighbours Regressor model using movies dataset.

CHAPTER-7

CONCLUSION

7.1 Conclusion

Prediction of movie success principally depends on numerous parameters, in our paper we've used some important parameters for delicacy and success prediction, besides this success also depends on some other factors i.e., Connection with the followership, Different Concept, Level of impact and numerous other effects have barred these parameters. We've done web scraping along with parameters like star cast, time, budget. Using these parameters delicacy is estimated using Random Forest Regression, K- Neighbors Regressor, Gaussian Process Regressor and prognosticated whether the movie will be Flop, megahit and Super megahit. We conclude that in this paper the model provides further delicacy than other models for predicting the movie. In moment's generation nearly every youth has their own account on each social media platform. Constantly these coffers are used for getting streamlined the information can be the justice score, or stock request, or about the launch of a new products etc. getting told by these coffers taking the star cast, directors, budget etc. As parameters will help chancing accurate in predicting the movie's success. Prediction will be done using different algorithms and also the delicacy will be analyzed. We can further incorporate our perpetration for prognosticating the success rate of web series and media.

7.2 LIMITATIONS AND FUTURE WORK:

Limitations:

- The success of the movie is often unpredictable and even well made movies with big budgets and star studded casts can flop at the box office.
- Predicting the success of the movie suffers from sampling bias, where the dataset used to make predictions may not be representative of the larger population, leading to inaccurate predictions.

Future Work:

- In Future, such system can be extended to recommender system by using some additional features in the form of hybrid techniques.
- Instead of just outputting a single prediction, this could be made as an interactive dashboard that allows users to input their own movie features and get a predicted success score.

BIBLIOGRAPHY

PAPER REFERENCES:

[1]. Partha Chakraborty,Md. Zahidur Rahman and Saifur Rahman,"Movie Success Prediction using Historical and Current Data Mining " ,International Journal of Computer Applications (0975 - 8887)Volume 178 - No.47, September 2019

[2]. Narayana Darapaneni, Christopher Bellarmine, Anwesh Reddy Paduri, Sujana, Entoori, Abir Kumar, SV Vybhav and Koushik Mondal."Movie Success Prediction Using ML," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 2020, pp. 0869-0874, doi: 10.1109/UEMCON51285.2020.9298145.

[3]. S. Sahu, R. Kumar, M. S. Pathan and J. Shafi et al., "Movie Popularity and Target Audience Prediction Using the Content-Based Recommender System," in IEEE Access, vol. 10, pp. 42044-42060, 2022, doi: 10.1109/ACCESS.2022.3168161.

[4]. Olubukola D.A., Stephen O.M,), Funmilayo A. Kasali , Ayokunle Omotunde ,Oyebola Akande , Oduroye Ayorinde , Wumi Ajayi and Yaw Mensah.(2021), Movie Success Prediction Using Data Mining. British Journal of Computer, Networking and Information Technology 4(2), 22-30. DOI: 10.52589/BJCNITCQOCIREC.

[5]. Dewan Muhammad Qaseem, Nashit Ali Waseem Akram, Aman Ullah, Kemal Polat (2022). Movie Success-Rate Prediction System through Optimal Sentiment Analysis. Journal of the Institute of Electronics and Computer, 4, 15-33. https://doi.org/10.33969/JIEC.2022.41002.

[6]. T. Sharma, R. Dichwalkar, S. Milkhe and K. Gawande, "Movie Buzz - Movie Success Prediction System Using Machine Learning Model," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 111-118, doi: 10.1109/ICISS49785.2020.9316087.

[7]. Hrithik Jain, Manohar Bhati, Sailee Zodape and Blessy Varghese, "Movie Success Prediction Using Data Mining", April 4th,2022. International Journal of Creative Research Thoughts(IJCRT) on an International Open Access, Peer- Reviewed, Refereed Journal, Volume 10, ISSN:2320-2882

[8]. Dipak Gaikar, Riddhi Solanki and Harshada Shinde, "Movie Success Prediction Using Popularity Factor From Social Media", International Research Journal Of Engineering and Technology (IRJET), Volume 6, Issue 4, April 2019.p-ISSN: 2395-0072, e-ISSN: 2395-0056.

[9]. M. Ahmed, M. Jahangir, H. Afzal, A. Majeed and I. Siddiqi, "Using Crowd-Source Based Features from Social Media and Conventional Features to Predict the Movies Popularity," in 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), Chengdu, China, 2015 pp. 273-278.doi: 10.1109/SmartCity.2015.83

[10]. Kumar, Saurabh. (2019). Movie Success Prediction using Data Mining For Data Mining and Business Intelligence(ITA5007) of Master of Computer Application School Of Information Technology and Engineering.

WEBSITE REFERENCES:

1. https://www.researchgate.net/publication/335878983_Movie_Success_Prediction_using_Historical_and_Current_Data_Mining
2. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9298145&isnumber=9297927
3. https://ieeexplore.ieee.org/abstract/document/9758691
4. https://abjournals.org/bjcnit/wpcontent/uploads/sites/11/journal/published_paper/volume-4/issue-2/BJCNIT_CQOCIREC.pdf
5. https://iecscience.org/jpapers/118
6. https://ieeexplore.ieee.org/abstract/document/9316087/keywords#keywords
7. https://ijcrt.org/papers/IJCRT2204495.pdf
8. https://www.academia.edu/44243744/Movie_Success_Prediction_Using_Popularity_Fa
9. https://doi.ieeecomputersociety.org/10.1109/SmartCity.2015.83
10. https://www.researchgate.net/publication/332396741_Movie_Success_Prediction_using_Data_Mining_For_Data_Mining_and_Business_IntelligenceITA5007_of_Master_of_Computer_Application_School_Of_Information_Technology_and_Engineering

APPENDICES

app.py :

```
import os

import pickle

import numpy as np

app=Flask(__name__)

@app.route('/',methods=["GET","POST"])

def index():

    return render_template('index.html')

@app.route("/predict",methods=["POST"])

def predict():

    gross=float(request.form["gross"])

    fblikes=float(request.form["fblikes"])

    movie_name = request.form["movie_name"]

    print("fblikes:",fblikes)

    print("gross:",gross)

    cur_dir=os.path.dirname(__file__)
k_Neighbors_model=pickle.load(open(os.path.join(cur_dir,'pickle_files/k_Neighbors_regressor_
model.sav'),'rb'))

    ip_arr = np.array([[fblikes,gross]])

   #list1=[James Cameron,Gore Verbinski      ,Sam Mendes  ,Christopher

Nolan,Andrew Stanton,Scott Smith,Benjamin Roberds,Daniel Hsia,Jon Gunn        ]

    #l1=list[]
```

```
    #pred=clf.predict(ip_arr)
    k_pred=k_Neighbors_model.predict(ip_arr)
    print("Prediction : ", k_pred)
    if ((k_pred) >=7.9):
        result="Super Hit"
    elif(((k_pred) >5) & ((k_pred)<8)):
        result=" Hit"
    else:
        result="Flop"
    #print("Prediction:" , pred)
    return render_template('result.html', result=result, movie_name=movie_name)
if __name__== '__main__':
    app.run()
```

index.html:

```
<!DOCTYPE html>
<html lang="en">
<head>
  <title>Movie Success Prediction</title>
  <meta charset="utf-8">
  <meta name="viewport" content="width=device-width, initial-scale=1">
 <linkhref="https://cdn.jsdelivr.net/npm/bootstrap@5.2.3/dist/css/bootstrap.min.css"
rel="stylesheet">
 <script src="https://cdn.jsdelivr.net/npm/bootstrap@5.2.3/dist/js/bootstrap.bundle.min.js">
</script>
 <div class="p-5 bg-primary text-white text-center">
 <h1>MOVIE SUCCESS PREDICTION</h1>
 <p>Lets find movie is success or failure!</p>
</div>
 <style>
 input[type=text]
  {
    padding:5px;
    font-size:18px;
    }
    input[type=submit]
    {
```

```
    padding:5px;
    font-size:18px;
    }
    .body
    {
     background-image:url('.\templates\im2.jpg');
     background-size:cover;
     width: 200px;
     height: 100vh;
    }
    .form-center
    {
    display:flex;
    justify-content: center;
    }
    </style>
    </head>
    <body>
<div class="form-center">
<form action="/predict" method="POST" >
<br>
<br><br>
<label> Content Rating</label><br>
```

```
<select id="cnt" name="cnt">

<option>----------</option>

<option>PG-13</option>

<option>TV-14</option>

<option>R</option>

<option>PG</option>

</select>

<br><br>

<label> Year</label><br>

<input type="text" id="year" name="year" required/>

<br><br>

<label> Movie Name</label><br>

 <select name="movie_name"  id="movie_name" required/>

<option>Select</option>

<option>Avatar</option>

<option>Pirates of the Caribbean:At World's End</option>

<option>Spectre</option>

<option>Battleship</option>

<option>The Dark Knight Rises</option>

<option>John carter</option>

<option>Signed Sealed Delivered</option>

<option>A plague so pleasant</option>

<option>My Date with Drew</option>
```

<option>Jason X</option>

<option>The Conjuring2</option>

<option>Love and Death on Long Island</option>

<option>Tangled</option>

<option>Avengers</option>

<option>Harry Potter and the Half Blood Prince</option>

<option>Pirates of the Carribbean</option>

<option>The Good Dinosaur</option>

<option>A Christmas Carol</option>

</select>

<label> Director Name</label>

<select name="director_name" id="director_name" >

<option>Select</option>

<option>James Cameron</option>

<option>Gore Verbinski</option>

<option>Sammendes</option>

<option>Peter Berg</option>

<option>Christopher Nola</option>

<option>Andrew Stanton</option>

<option>Scott Smith</option>

<option>Benjamin Roberd</option>

<option>Jon Gunn</option>

<option>James Isaac</option>

<option>James Wan</option>

<option>Richard Kwietniowski</option>

<option>Nathan Greno</option>

<option>Joss Whedon</option>

<option>David Yates</option>

<option>Gore Verbinski</option>

<option>RobertZemeckis</option>

</select>

<label> Actor Name</label>

<select name="actor_name"/>

<option>Select</option>

<option>CCH pounder</option>

<option>Johnny Deep</option>

<option>Christopher Waltz</option>

<option>Liam Neeson</option>

<option>Tom Hardy</option>

<option>Daryl Sabara</option>

<option>Eric Mabius</option>

<option>Eva Boehnke</option>

<option>John August</option>

<option>Peter Mensah</option>

<option>Javier Bote</option>

<option>Jason Priestley</option>

<option>Brad Garrett</option>

<option>Chris Hemsworth</option>

<option>Alan Rickman</option>

<option>Johny Depp</option>

<option>Robin Wright</option>

</select>

<label> Genres</label>

<select id="genres" name="genres"/>

<option>Select</option>

<option>Action</option>

<option>Comedy</option>

<option>Thriller</option>

<option>Romance</option>

<option>Fantasy</option>

<option>Horror</option>

<option>Adventure</option>

<option>Sci-Fi</option>

<option>Drama</option>

<option>Documentary</option>

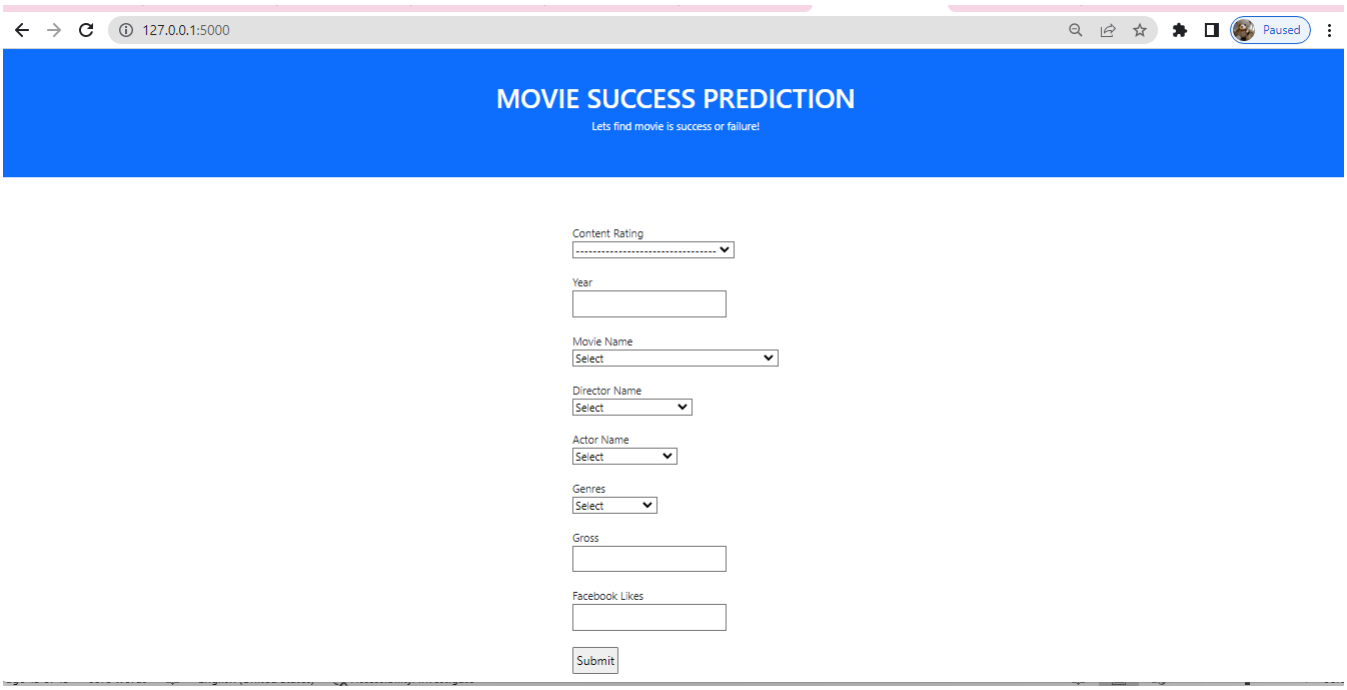
<option>Crime</option>

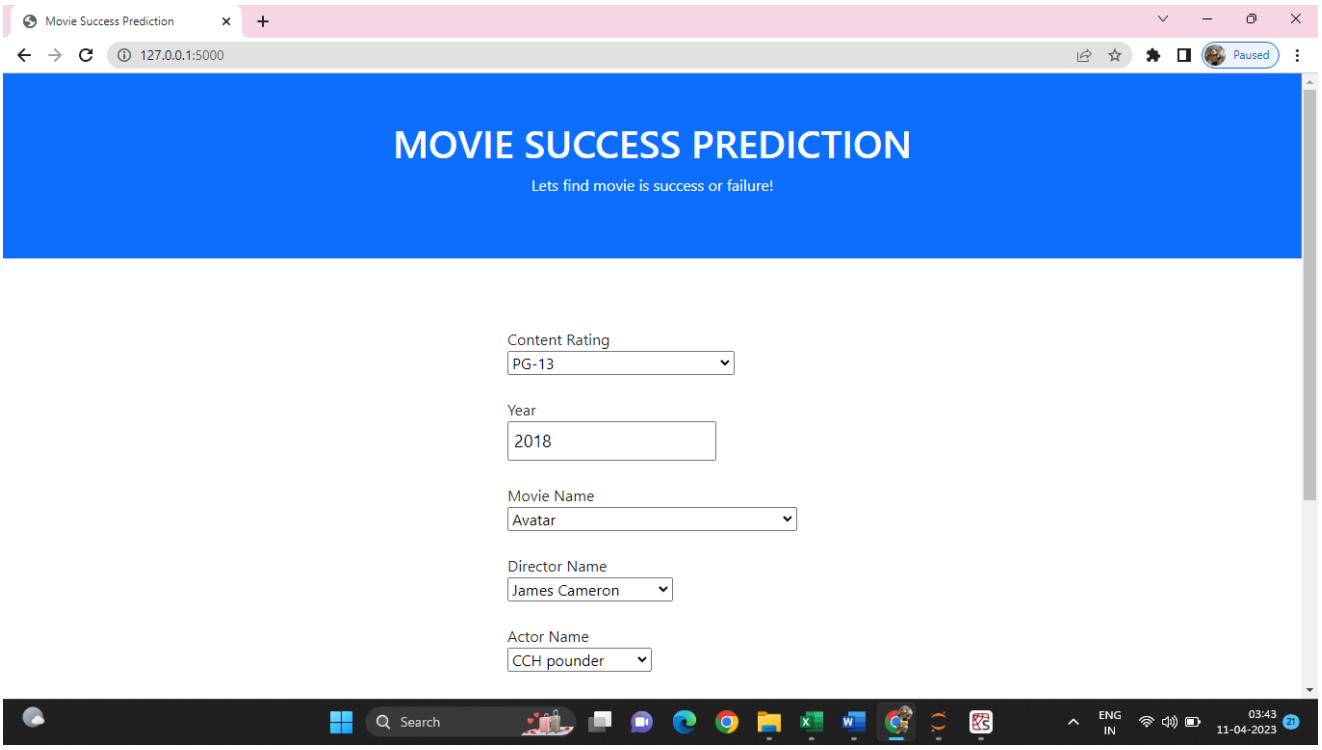
```
</select>

<br><br>

<label> Gross</label><br>

<input type="text" name="gross" required/>

<br><br>

<label> Facebook Likes</label><br>

<input type="text" name="fblikes" required/>

<br>

<br>

<input type="Submit" name="predict" value="Submit"/>

<br>

</form>

</div>

</body>

</html>
```

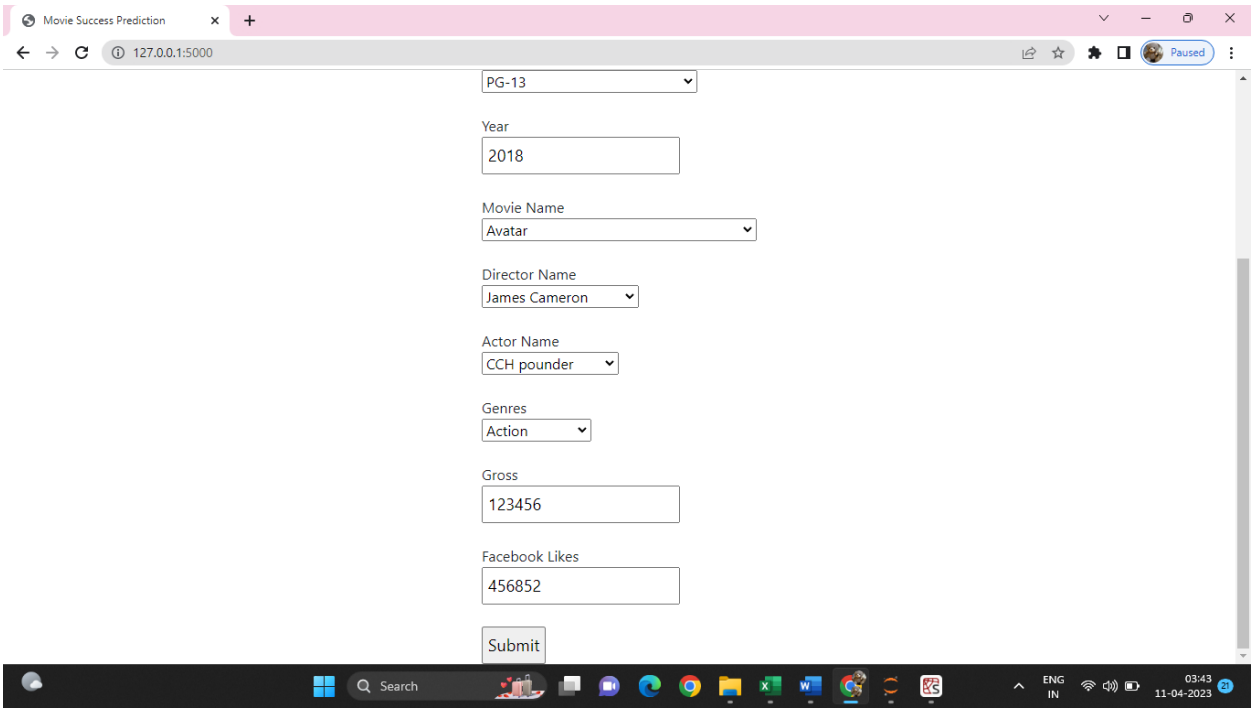
result.html:

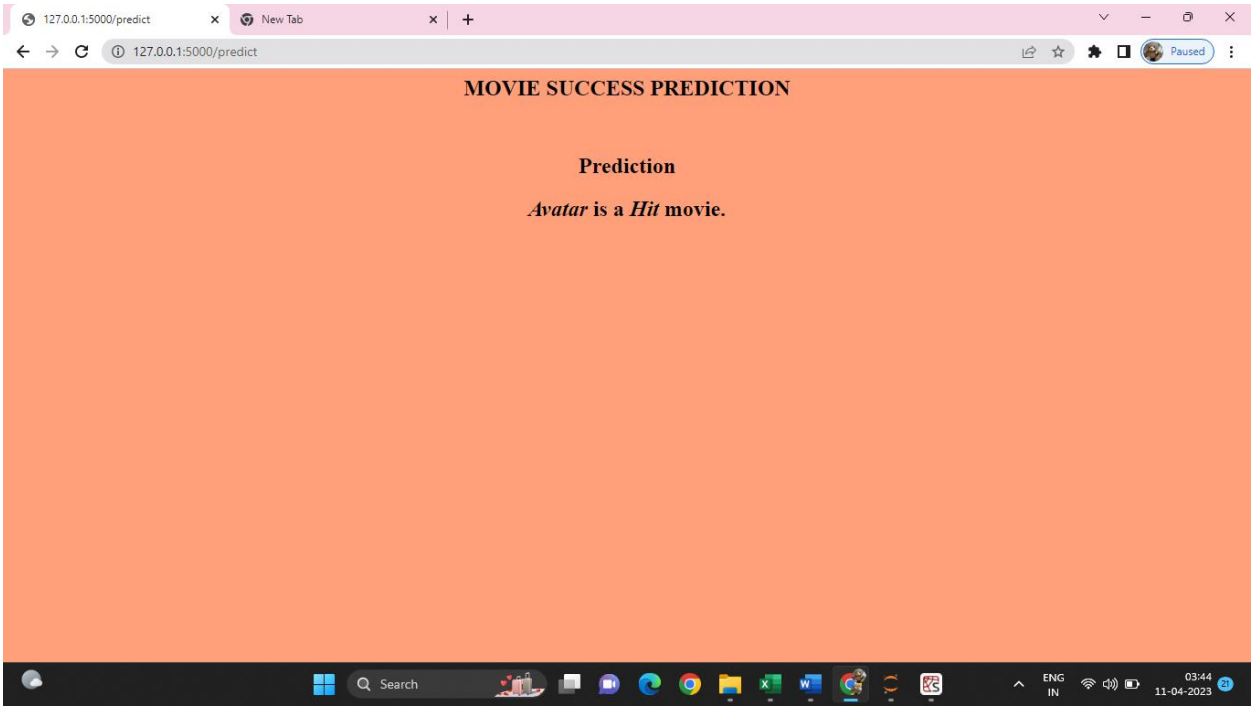
```
<html>
<head>
<style>
body
{
background:#FFA07A
};
</style>
</head>
<body>
<h2 align="center">MOVIE SUCCESS PREDICTION</h2>
<br>
<h2 align ="center">Prediction</h2>
<h2 align="center">
<b><i>{{ movie_name }}</i></b> is a <b><i>{{ result }} </i></b>movie.
</h2>
</body>
</html>
```

Output:


127.0.0.1:5000
Paused
MOVIE SUCCESS PREDICTION
Lets find movie is success or failure!
Content Rating
Year
Movie Name
Select
Director Name
Select
Actor Name
Select
Genres
Select
Gross
Facebook Likes
Submit


Movie Success Prediction
127.0.0.1:5000
Paused
MOVIE SUCCESS PREDICTION
Lets find movie is success or failure!
Content Rating
PG-13
Year
2018
Movie Name
Avatar
Director Name
James Cameron
Actor Name
CCH pounder
Search
ENG
IN
03:43
11-04-2023


Movie Success Prediction
127.0.0.1:5000
Paused
PG-13
Year
2018
Movie Name
Avatar
Director Name
James Cameron
Actor Name
CCH pounder
Genres
Action
Gross
123456
Facebook Likes
456852
Submit
Search
ENG
IN
03:43
11-04-2023


127.0.0.1:5000/predict
New Tab
127.0.0.1:5000/predict
Paused
MOVIE SUCCESS PREDICTION
Prediction
Avatar is a Hit movie.
Search
ENG
IN
03:44
11-04-2023