**Orchid International College**

**(Affiliated to Tribhuwan University)**

**Bijayachowk, Gaushala**



**A Final Year Project Report**

**On**

**MOVIE SUCCESS PREDICTION SYSTEM USING K-MEANS CLUSTERING ALGORITHM**

**For the partial fulfillment of Bachelor of Science in**

**Computer Science and Information Technology awarded by**

**Tribhuvan University**

**Under the Supervision of**

**Mr. Dhiraj Kumar Jha**

**Lecturer**

**Orchid International College**

**Submitted To:**

**Department of Computer Science and Information Technology**

**Orchid International College**

**Submitted By:**

**Kabita Rawal (8066/072)**

**Pratikshya Kaphle (8078/072)**

**August, 2019**



# **SUPERVISOR’S RECOMMENDATION LETTER**

I hereby recommend that the report prepared under my supervision by Kabita Rawal (TU Exam Roll No. 8066/072) and Pratikshya Kaphle (TU Exam Roll No. 8078/072) entitled “**MOVIE SUCCESS PREDICTION SYSTEM USING K-MEANS CLUSTERING ALGORITHM**” in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Information Technology be processed for evaluation.

**……………..…….**

**Dhiraj Kumar Jha**

Project Coordinator, Department of CSIT

Orchid International College

Bijayachowk, Gaushala

******

# **CERTIFICATE OF APPROVAL**

This is to certify that this project prepared by Kabita Rawal (TU Exam Roll No. 8066/072) and Pratikshya Kaphle (TU Exam Roll No. 8078/072) entitled “**MOVIE SUCCESS PREDICTION SYSTEM USING K-MEANS CLUSTERING ALGORITHM**” in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Information Technology has been well studied. In our opinion, it is satisfactory in the scope and quality as a project for the required degree.

|  |  |
| --- | --- |
| ------------------------------  **Dhiraj Kumar Jha**  Project Coordinator,  Department of Computer Science and IT  Orchid International College  Bijayachowk, Gaushala | -------------------------  **Bijay Mishra**  Program Coordinator,  Department of Computer Science and IT  Orchid International College  Bijayachowk, Gaushala |
| -------------------------  **External Examiner**  Central Department of Computer Science and IT  Tribhuvan University  Kirtipur, Nepal | |

# **ACKNOWLEDGEMENT**

It is a great pleasure to have the opportunity to extend our heartfelt gratitude to everyone who helped us throughout the course of this project. We would like to express our profound gratitude to our supervisor **Mr. Dhiraj Kumar Jha** for his guidance and support in each step for the progress of this project. His continuous inspiration and valuable suggestions has made us complete this project and achieve its target.

We would like to express our sincere gratitude towards the **Department of Computer Science and Information Technology, Orchid International College,** for providing us with the wonderful opportunity, encouragement and environment to explore our skills and gain learning experience through this major project.

We would also like to express our deepest appreciation to **Mr. Bijay Mishra,** program coordinator, Orchid International College, for his constant motivation, support and for providing us with a suitable working environment.

Last but not the least; we would like to thank all the faculty members and friends who has direct and indirect contribution to our project. Any kind of suggestions or criticism will be greatly appreciated and acknowledged.

Sincerely,

Kabita Rawal (T.U. Roll No. 8066/072)

Pratikshya Kaphle (T.U. Roll No. 8078/072)

# **ABSTRACT**

In real world prediction models and mechanisms can be used to predict the success of a movie. The proposed work aims to develop a system based upon data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty.

In this paper we developed a mathematical model for predicting the success class such as flop, hit, and average of the movies. For doing this we have to develop a methodology in which the historical data of each component such as actor, director and genre that influences the success or failure of a movie is given is due to weightage and then based on multiple thresholds calculated on the basis of descriptive statistics of dataset of each component it is given class flop, hit, average label. Based on the weightage of historical data of each film crew the movie will be labeled as hit, flop or average. This system helps to find out whether the movie is average, hit, flop on the basis of historical data of actor, director and genre data of the new movie.

The factors such as actor, director, and genre historical data of each component are calculated and movie success is predicted. This mathematical model helps to find out the review of the new movie. Due to this model, user can easily decide whether to book ticket in advance or not.

***Keywords:*** *Prediction, Data Mining, Classification, K-means algorithm, IMDb, Kaggle, Movie Dataset*

**Table of Contents**

[**SUPERVISOR’S RECOMMENDATION LETTER** ii](#_Toc17111683)

[**CERTIFICATE OF APPROVAL** iii](#_Toc17111684)

[**ACKNOWLEDGEMENT** iv](#_Toc17111685)

[**ABSTRACT** v](#_Toc17111686)

[**LIST OF FIGURES** viii](#_Toc17111687)

[**LIST OF TABLES** ix](#_Toc17111688)

[**ABBREVIATIONS** x](#_Toc17111689)

[**CHAPTER 1- INTRODUCTION** 1](#_Toc17111690)

[**1.1. Project Introduction** 1](#_Toc17111691)

[**1.2. Problem Statement** 2](#_Toc17111692)

[**1.3. Objectives** 2](#_Toc17111693)

[**1.4. Scope** 2](#_Toc17111694)

[**1.5. Report Organization** 3](#_Toc17111695)

[**CHAPTER 2- REQUIREMENT ANALYSIS** 4](#_Toc17111696)

[**2.1.** **Literature Review** 4](#_Toc17111697)

[**2.2. Requirement Analysis** 5](#_Toc17111698)

[**2.2.1. Functional Requirements** 5](#_Toc17111699)

[**2.2.2. Non Functional Requirements** 10](#_Toc17111700)

[**2.3. Feasibility Analysis** 10](#_Toc17111701)

[**CHAPTER 3- SYSTEM ANALYSIS** 12](#_Toc17111702)

[**3.1. Structuring System Requirements** 12](#_Toc17111703)

[**3.1.1. Data Modeling** 12](#_Toc17111704)

[**3.1.2. Process Modeling** 14](#_Toc17111705)

[**CHAPTER 4- SYSTEM DESIGN** 17](#_Toc17111706)

[**4.1. Database Schema** 17](#_Toc17111707)

[**4.2. Activity Diagram** 18](#_Toc17111708)

[**4.3. Sequence Diagram** 19](#_Toc17111709)

[**4.4. Input Output Design** 19](#_Toc17111710)

[**CHAPTER 5- IMPLEMENTATION** 21](#_Toc17111711)

[**5.1. Tools Used** 21](#_Toc17111712)

[**5.2. Methodology** 22](#_Toc17111713)

[**5.2.1. Data Collection** 23](#_Toc17111714)

[**5.2.2. Data Cleaning and Integration** 25](#_Toc17111715)

[**5.2.3. Data Transfer** 27](#_Toc17111716)

[**5.2.4. Data Analysis and Prediction** 28](#_Toc17111717)

[**CHAPTER 6- SYSTEM TESTING AND ANALYSIS** 34](#_Toc17111718)

[**6.1. Unit Testing** 34](#_Toc17111719)

[**6.2. Integration Testing** 36](#_Toc17111720)

[**6.3. Result Analysis** 37](#_Toc17111721)

[**CHAPTER 7- CONCLUSION** 41](#_Toc17111722)

[**7.1. Conclusion** 41](#_Toc17111723)

[**REFERENCES** 42](#_Toc17111724)

[**APPENDIX** 43](#_Toc17111725)

# **LIST OF FIGURES**

[**Figure 2‑1: Use Case Diagram for Movie Success Prediction System** 6](#_Toc17111972)

[**Figure 3‑1: ER Diagram for Movie Success Prediction System** 12](#_Toc17111973)

[**Figure 3‑2: Class Diagram for Movie Success Prediction System** 13](#_Toc17111974)

[**Figure 3‑3: Context Diagram for Movie Success Prediction System** 14](#_Toc17111975)

[**Figure 3‑4: Level 0 DFD for Movie Success Prediction System** 15](#_Toc17111976)

[**Figure 3‑5: Level 1 DFD for process 3.0 of level 0 DFD** 16](#_Toc17111977)

[**Figure 4‑1: Database Schema for Movie Success Prediction System** 17](#_Toc17111978)

[**Figure 4‑2: Activity Diagram for Predicting Result** 18](#_Toc17111979)

[**Figure 4‑3: Sequence Diagram of User in the system** 19](#_Toc17111980)

[**Figure 4‑4: Input Design of User in System** 20](#_Toc17111981)

[**Figure 4‑5: Output Design of User in System** 20](#_Toc17111982)

[**Figure 5‑1: Architecture of Prediction Model** 22](#_Toc17111983)

[**Figure 5‑2: Flowchart of K-means clustering algorithm** 33](#_Toc17111984)

[**Figure 6‑1: Test Result of Movie: Devdas** 36](#_Toc17111985)

[**Figure 6‑2: Test Result of Movie: Kalank** 37](#_Toc17111986)

[**Figure 6‑3: Pie-chart of actual and predicted movie data categories** 40](#_Toc17111987)

# **LIST OF TABLES**

[**Table 2‑1: UC 1 Registration** 6](#_Toc17103504)

[**Table 2‑2: UC 2 Admin login** 7](#_Toc17103505)

[**Table 2‑3: UC 3 User login** 7](#_Toc17103506)

[**Table 2‑4: UC 4 Manage movie training dataset** 7](#_Toc17103507)

[**Table 2‑5: UC 5 Test result** 8](#_Toc17103508)

[**Table 2‑6: UC 5 Review Movie details** 8](#_Toc17103509)

[**Table 2‑7: UC 7 Search the movie** 8](#_Toc17103510)

[**Table 2‑8: UC 8 Post Comments** 9](#_Toc17103511)

[**Table 2‑9: UC 9 Predict Movie** 9](#_Toc17103512)

[**Table 5‑1: Bollywood Movie Detail** 24](#_Toc17103513)

[**Table 5‑2: Bollywood Actor Rating** 25](#_Toc17103514)

[**Table 5‑3: Bollywood Director Rating** 25](#_Toc17103515)

[**Table 5‑4: Filtered Bollywood Movie Detail** 26](#_Toc17103516)

[**Table 5‑5: Filtered Bollywood Actor Rating** 26](#_Toc17103517)

[**Table 5‑6: Filtered Bollywood Director Rating** 26](#_Toc17103518)

[**Table 5‑7: Filtered Bollywood Genre Rating** 26](#_Toc17103519)

[**Table 5‑8: Movie-Actor mapping** 27](#_Toc17103520)

[**Table 5‑9: Movie-Director mapping** 27](#_Toc17103521)

[**Table 5‑10: Movie-Genre mapping** 27](#_Toc17103522)

[**Table 5‑11: Training dataset** 27](#_Toc17103523)

[**Table 5‑12: Testing dataset** 28](#_Toc17103524)

[**Table 5‑13: Consideration for Class** 28](#_Toc17103525)

[**Table 6‑1: Unit Testing of Movie Prediction System** 34](#_Toc17103526)

[**Table 6‑2: Integration Testing of Movie Prediction System** 36](#_Toc17103527)

[**Table 6‑3: Confusion Matrix** 38](#_Toc17103528)

[**Table 6‑4: F-measure of different categories** 39](#_Toc17103529)

[**Table 6‑5: Summary of result analysis** 40](#_Toc17103530)

# **ABBREVIATIONS**

|  |  |
| --- | --- |
| **CSS** | Cascading Style Sheets |
| **CSV** | Comma-Separated Values |
| **DFD** | Data Flow Diagram |
| **ECMA** | European Computer Manufacturers Association |
| **ERD** | Entity Relationship Diagram |
| **HTML** | Hypertext Markup Language |
| **IMDb** | Internet Movie Database |
| **OS** | Operating System |
| **PHP** | Hypertext Pre Processor |
| **RF** | Random Forests |
| **SQL** | Structured Query Language |
| **SVM** | Support Vector Machine |
| **UC** | Use Case |
| **UI** | User Interface |
| **UML** | Unified Modeling Language |

# **INTRODUCTION**

## **1.1. Project Introduction**

The Indian movie industry produces the maximum number of movies per year at 1000/year, higher than any other country’s movie industry. The IMDb is an excellent resource to find detailed information about almost any film ever made. It contains a vast amount of data, which undoubtedly contains much valuable information about general trends in films. Data mining techniques allows us to predict the success of a future film given select information about the film before its release. So we will get the data from Kaggle, which is one of the site where historical data about movie is stored and predict the movie success by applying the mathematical concept. The main challenge is to collect them all and apply cleaning and integration on them to obtain the relevant data. Once this is done, we will use K-means clustering algorithm to classify the general rating of upcoming films based on the historical data collected and predict whether the movie will be successful or not.

Given the low success rate of movies, models and mechanisms can be used to predict the success of a movie. It will help the business significantly. Various stakeholders such as actors, producers, directors etc. can use these predictions to make more informed decisions. They can make the decision before the movie release. Historical data of each component such as actor, actress, director and genre that influences the success or failure of a movie is given due to its weightage.

Movie prediction aims to develop a model based upon the data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. The system is used to predict the past as well as the future of movie for the purpose of business certainty or simply a theoretical condition in which decision making (the success of the movie) is without risk, because the decision maker (movie makers and stake holders) has all the information about the exact outcome of the decision, before he or she makes the decision (release of the movie).

## **1.2. Problem Statement**

The audience expectation from the movie may not be fulfilled without proper analysis of the movie before buying tickets for the movie they are willing to watch. Moreover, movies involves huge investment. Before new movies are being produced, every stakeholder is interested in the monetary success of the intended movie. In order to predict the success, costly methods are being applied, such as market investigations or analyses. India, is a movie fanatic nation, but there is no composite movie database or a credible source for movie-ratings.

In order to solve such problems, the movie success prediction system implements a data mining algorithm: K-means clustering. This algorithms was used in the system to predict the success or failure of movie. The benefit of data mining to analyze large datasets and can also be transferred to the stated problem of predicting a movie’s success.

## **1.3. Objectives**

Our application focuses on making people easier and less of a hassle. The major objectives are:

* To extract patterns and trends which can be beneficial in predicting movies success.
* To predict the movie using K-means clustering algorithm.
* To provide more precise prediction of success, hence providing confidence to stakeholders in their investments.

## **1.4. Scope**

Movie Prediction System can be used to predict movies before their release so that people can get idea about the movie whether or not to waste their precious time watching the movie, i.e. whether it is worth watching the movie. It will save their money as well as valuable time. The Movie Prediction System provides an attractive user interface for the end users in which they can choose the name of the upcoming movie and get the prediction of the movie just with a click of the “PredictMovie” Button. This system only predicts the movie based on the variables genre, actor and director and does not take into considerations of other variables. This system can only predict if genre, actor and director data are already available in the database. For the new variable values, the system fails to predict correctly. Because the cast name which is not available in database will give 0 rating of respective cast. For this reason K-means algorithms fails to give correct output.

## **1.5. Report Organization**

This report is organized into 6 chapters as follows:

**Chapter 1:**

The project is introduced in detail with its objectives and scope.

**Chapter 2:**

It contains the functional and non-functional requirements of the project. Analysis and evaluation of project is done by feasibility analysis.

**Chapter 3:**

Data modeling and process modeling of the project is done to analyze the data and working mechanism of the system in detail.

**Chapter 4:**

Explains the methods and tools used to implement the project.

**Chapter 5:**

Different Test Cases used during testing of each component of the system.

**Chapter 6:**

It contains conclusion and recommendations based on the project.

# **REQUIREMENT ANALYSIS**

## **2.1.** **Literature Review**

With over two million Audience a day and films exported to over 100 countries, the impact of Bollywood film industry is formidable. From the first Indian film, India produced over 15000 feature films [1]. Since then it has produced, at least another 15000 at a rate of more than 1000 films a year in 26 languages. Literature survey has revealed many studies which have attempted to predict the success of movies. While one study uses Bayesian belief network to predict the success, the other one uses neural network for the same. Bayesian belief network for predicting box office performance concluded that Bayesian belief networks were better in predicting the success as compared to neural networks [2].

Machine learning has also been used for predicting movie success by using algorithms like RF and SVM. Although the use of RF and SVM within the movie domain seems to be fairly limited, the two algorithms have been applied and evaluated in many applications for the purpose of regression as well as classification. Within recent study [3] have surveyed a number of large as well as small scale comparisons on data mining and machine learning, all of which include the RF algorithm, specifically issuing its prediction performance in comparison to other algorithms as well as the use of the variable importance estimates available from RF. Among the previous applications and algorithm comparisons, there are several large scale studies, evaluating RF and SVM among other algorithms over a number of 33 and 22 datasets respectively.

Similarly, a method of selectively choosing the relevant data sets from IMDb over a total of 49 files was proposed [4]. Java application has been used for cleaning and integration of the files and then storing them in databases. Finally, Microsoft Envisioner and Microsoft Excel has been used for the actual data mining process.

This paper entitled “Sentiment analysis of Indian movie review with various feature selection techniques” [5] details the analysis of the Internet Movie Database (IMDb), a free, user-maintained, online resource of production details for over 390,000 movies, television series and video games, which contains information such as title, genre, box-office taking, cast credits and user's ratings.

One of the paper, Feature Level Sentiment Analysis on Movie Reviews ACM Computing Surveys, [6] has used simple Machine learning techniques including twitter sentiment analysis to generate an overall review based on interactive inputs from others.

## **2.2. Requirement Analysis**

Requirements analysis encompasses those tasks that go into determining the needs or conditions to meet for a new or altered product or project. Requirement analysis is mainly categorized into two types:

### **2.2.1. Functional Requirements**

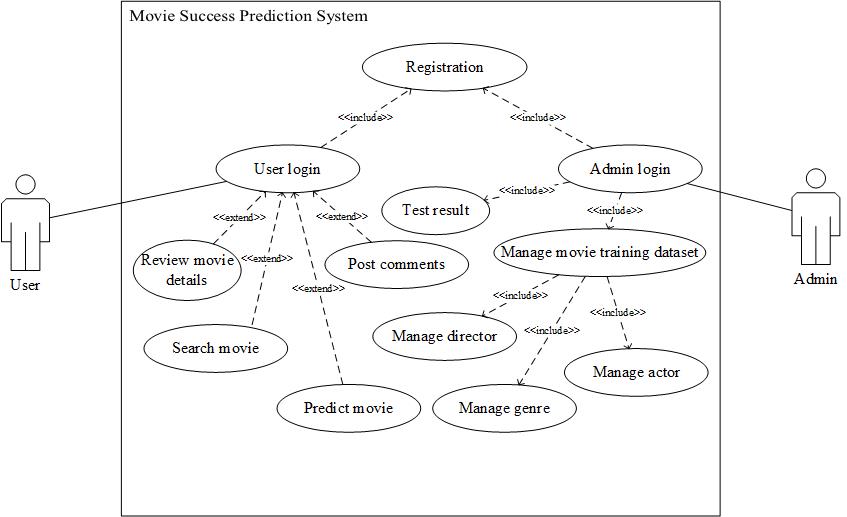
The functional requirements for a system describe what the system should do. Those requirements depend on the type of software being developed, the expected users of the software. These are statement of services the system should provide, how the system should react to particular inputs and how the system should behave in particular situation.

Movie success prediction system has a solo thing to achieve that is to predict after the user has provided data from which data is to be extracted and then summarized. After the user has provided all required inputs the system provides a result. If the user has given any malicious input, the system returns unexpected output which may cause bad reputation of the crew members.

The functional requirements of this system are as follows:

* Allow user to input data
* Allow user to search the movie
* Allow user to give feedback of the movie
* Allow user to rate the movie
* Allow user to view the details of the movie
* Allow user to login in or sign up in the system
* Allow user and admin to see the prediction result
* Allow the admin to manage the users
* Allow the admin to add training data set

The following Use Case Diagram describes the major actions of the system and interaction between actors (user, admin) and the system in our application:



**Figure 2‑1: Use Case Diagram for Movie Success Prediction System**

The use case diagram of the movie success prediction system is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. This diagram shows that the user can search and view the details of the movies. The user can also see the prediction result. In addition the user can also rate the movie.

**Use-case Description**

**Table 2‑1: UC 1 Registration**

|  |  |
| --- | --- |
| Primary Actor | Admin and User |
| Secondary Actor | NONE |
| Description | Admin as well as User registers his/her account in the system in order to login into the system |
| Pre-condition | Both admin and user have to fill up the requirement field to register |
| Post-condition | Both will be able to login to the system |
| Failure Scenario | Missing required entry of registration info |

**Table 2‑2: UC 2 Admin login**

|  |  |
| --- | --- |
| Primary Actor | Admin |
| Secondary Actor | NONE |
| Description | Admin login to the system |
| Pre-condition | Admin has to register in the system in order to proceed further operation |
| Post-condition | Admin will be able to manage the crew members data |
| Failure Scenario | Invalid credentials |

**Table 2‑3: UC 3 User login**

|  |  |
| --- | --- |
| Primary Actor | User |
| Secondary Actor | NONE |
| Description | User login to the system |
| Pre-condition | User has to register in the system in order to proceed further operation |
| Post-condition | User will be able to search, rate and give feedback of the movies as well as predict the movie |
| Failure Scenario | Invalid credentials |

**Table 2‑4: UC 4 Manage movie training dataset**

|  |  |
| --- | --- |
| Primary Actor | Admin |
| Secondary Actor | NONE |
| Description | Admin manage the dataset |
| Pre-condition | Admin has to register in the system and have precise data to insert into the database |
| Post-condition | Admin will be able to insert the data which are not in database |
| Failure Scenario | Database connection inactive |

**Table 2‑5: UC 5 Test result**

|  |  |
| --- | --- |
| Primary Actor | Admin |
| Secondary Actor | NONE |
| Description | Admin test the result |
| Pre-condition | Admin have to register in the system and have data which are inserted in database |
| Post-condition | Admin will be able to test the result which improve efficiency |
| Failure Scenario | If data are not in database, system will fail to give accurate result |

**Table 2‑6: UC 5 Review Movie details**

|  |  |
| --- | --- |
| Primary Actor | User |
| Secondary Actor | NONE |
| Description | User review the movie details |
| Pre-condition | User have to register in the system |
| Post-condition | User will be able to review the details of the movie such as movie’s crew member |
| Failure Scenario | If no movie then user will be not able to review the movie |

**Table 2‑7: UC 7 Search the movie**

|  |  |
| --- | --- |
| Primary Actor | User |
| Secondary Actor | NONE |
| Description | User search the movie |
| Pre-condition | User has to register in the system and searched movie should be in database |
| Post-condition | User will get the movie info |
| Failure Scenario | If there is no movie then user will be not able to search the movie |

**Table 2‑8: UC 8 Post Comments**

|  |  |
| --- | --- |
| Primary Actor | User |
| Secondary Actor | NONE |
| Description | User give feedback of the movie |
| Pre-condition | User have to register in the system |
| Post-condition | User can read post |
| Failure Scenario | No database connection and user not logged in the system |

**Table 2‑9: UC 9 Predict Movie**

|  |  |
| --- | --- |
| Primary Actor | User |
| Secondary Actor | NONE |
| Description | User predict the movie |
| Pre-condition | User have to register in the system and have crew member data |
| Post-condition | User will be able to predict the movie by providing details of the movie |
| Failure Scenario | If data are not available in database, system will fail to predict the movie |

### **2.2.2. Non Functional Requirements**

Non-functional requirements are requirements that are not directly concerned with the specified function delivered by the system. They may relate to emergent system properties such as reliability, response time and store occupancy. Some of the non-functional requirements related with this system are hereby below:

* Reliability:

Reliability based on this system defines the evaluation result of the system, correct prediction of the movie. The system must be able to generate output for all possible inputs.

* Ease of Use:

The system is simple, user friendly and graphics user interface implemented is easy to navigate so that anyone can use this system without any difficulties.

* Scalability:

The movie success predication system must be able to predict not only of Hindi movie but also English, Nepali simply by adding required feature set.

## **2.3. Feasibility Analysis**

Before starting the project, feasibility study is carried out to measure the viable of the system. Feasibility study is necessary to determine if creating a new or improved system is friendly with the cost, benefits, operation, technology and time. Following are the feasibility that is concerned in this project:

* Technical Feasibility:

Technical feasibility is one of the first studies that must be conducted after the project has been identified. Technical feasibility study includes the hardware and software devices. The required technologies (PHP, My-SQL) existed.

* Operational Feasibility:

The purposed system is compatible and easy to use. The proper guidance from the supervisor existed. Any user with basic knowledge of computer and internet can access it.

* Economic Feasibility:

The required resources were freely available for the purposed system. It can be assured that the system is economically feasible.

# 

# **SYSTEM ANALYSIS**

## **3.1. Structuring System Requirements**

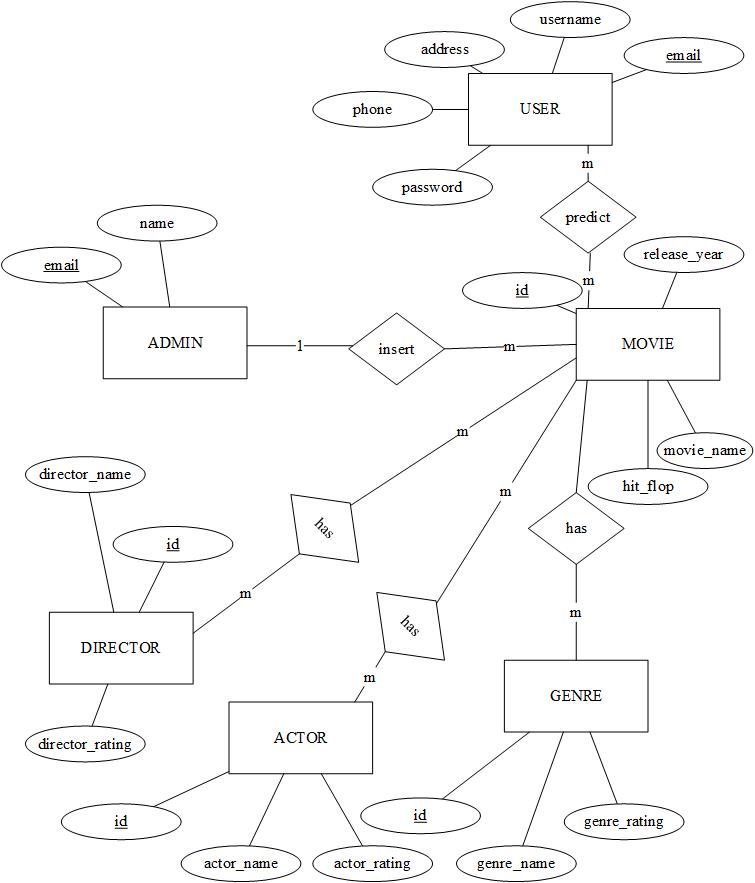
The field of system analysis relates closely to requirements analysis. It is also an explicit formal inquiry carried out to help a decision maker identify a better course of action. It contains the unpacking of the system requirements from data modeling and process modeling of the system.

### **3.1.1. Data Modeling**

Data modeling is a process used to define and analyze data requirements needed to support the business process within the scope of corresponding information systems in organizations.

**ER – Diagram**

An ER model describes inter-related things of interest in a specific domain of knowledge. The following ER model shows the entities, their attributes and relationships between them in our application.

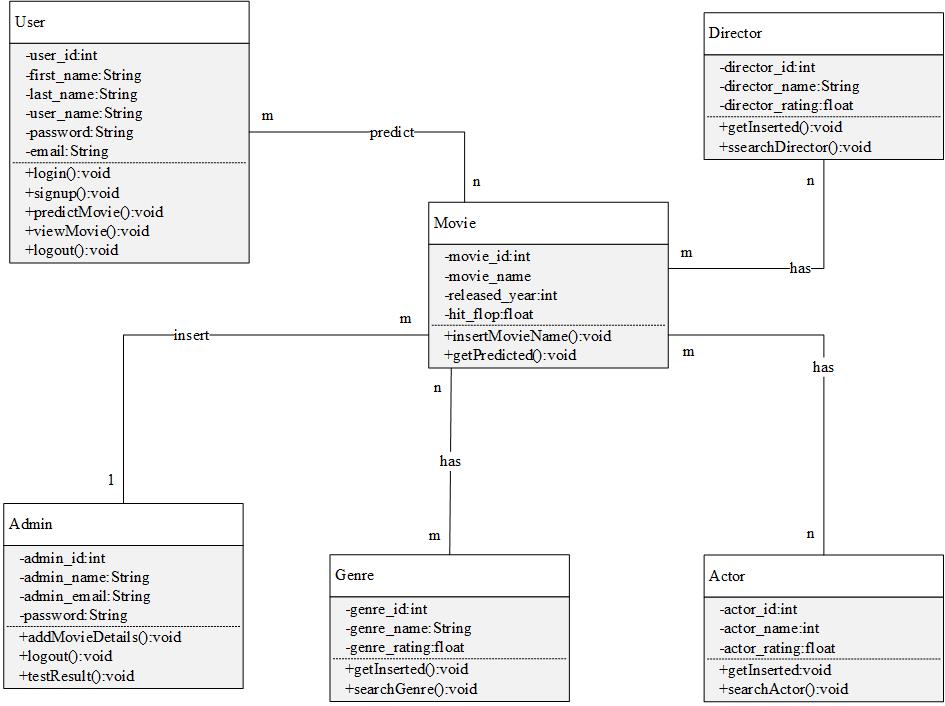


**Figure 3‑1: ER Diagram for Movie Success Prediction System**

The above ER diagram of the purposed system is simple. There are three main entities namely user, admin and movie. Many users can view multiple movies so there existed many to many relationship between them. Likewise the user entity has several attributes namely name, username, email, etc. In addition the admin entity has attributes namely name and email and has one to many relationship with movie. Movie has its own attribute which include movie name, release year and movie rating. Movie has several other entities namely actor, director and genre which include several attributes respectively and has many to many relationship with movie.

**Class Diagram**

The following class diagram is an illustration of the relationships and source code dependencies among classes in our system. Each class has three sections: name, attributes and operations.



**Figure 3‑2: Class Diagram for Movie Success Prediction System**

### **3.1.2. Process Modeling**

Process modeling is a technique for organizing and documenting the structure and flow of data through a system's processes and/or the logic, policies, and procedures to be implemented by a system's processes.

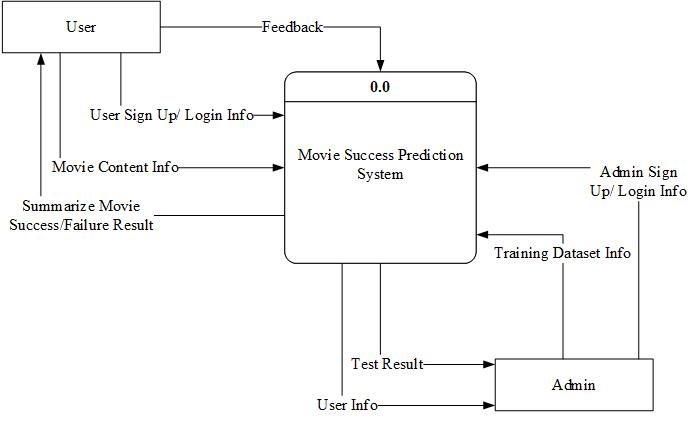
**Data Flow Diagram**

Data flow diagram (DFD) is an analysis tool to represent the flow of data through an information system. The data flow diagram of our system is divided into context diagram and level 0 DFD.

**Context Diagram**

A context diagram defines the boundary between the system, or part of a system, and its environment, showing the entities that interact with it. This diagram is a high level view of a system. It is similar to a block diagram.

The context diagram of our system is shown as below:

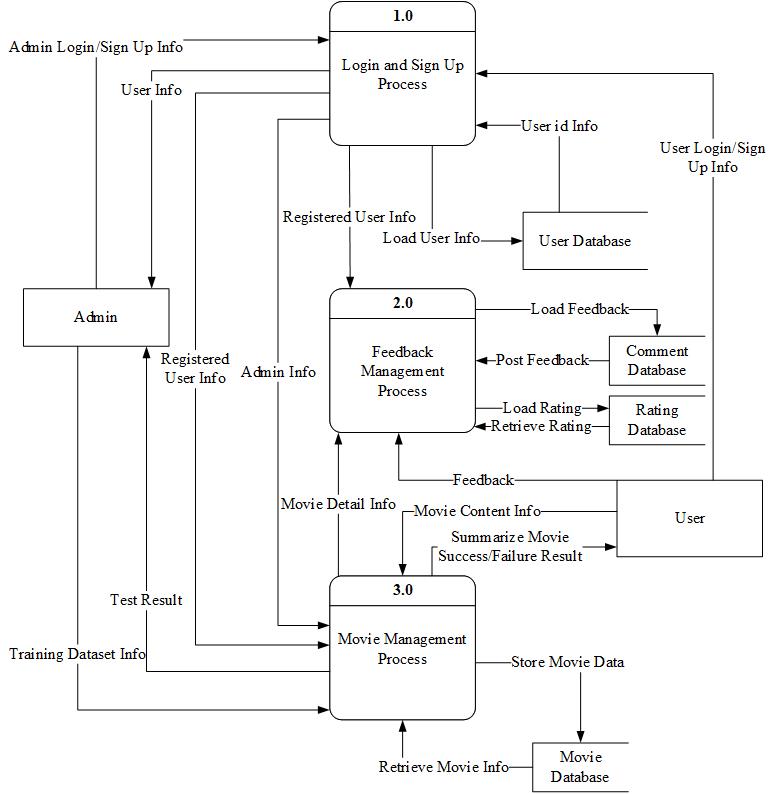


**Figure 3‑3: Context Diagram for Movie Success Prediction System**

The above context diagram shows that the purposed system has two distinct entities namely user and admin with a movie success prediction system. All of these entities interact with each other in several ways as shown in the figure above.

**Level 0 DFD**

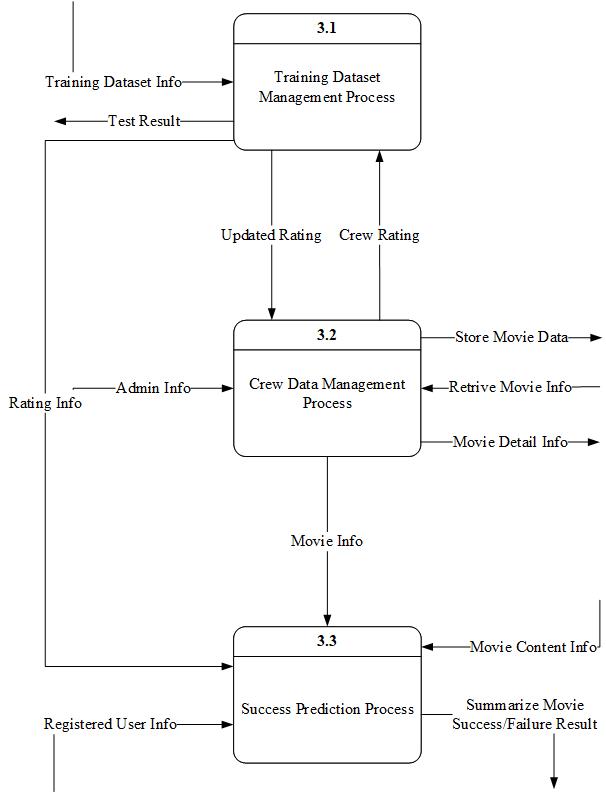
The level 0 DFD diagram of the proposed system shows the data flow in the movie prediction system. The admin can manage the user info and update the user database accordingly. Likewise the user can provide login info which is authenticated by accessing the user database. The user is also allowed to provide movie’s crew details as an input to the prediction system. The K-mean process can generate the result as per the input provided by the user in the system. The level 0 DFD of the above context diagram is shown as below:



**Figure 3‑4: Level 0 DFD for Movie Success Prediction System**

**Level 1 DFD**

In the level 1 DFD for process 3.0 of level 0 DFD, there are three major processes namely training dataset management process, crew data management process and success prediction process. The user can provide movie info in success prediction process and then get the result whether movie is hit or flop. The level 1 DFD of process 3.0 is shown as below:

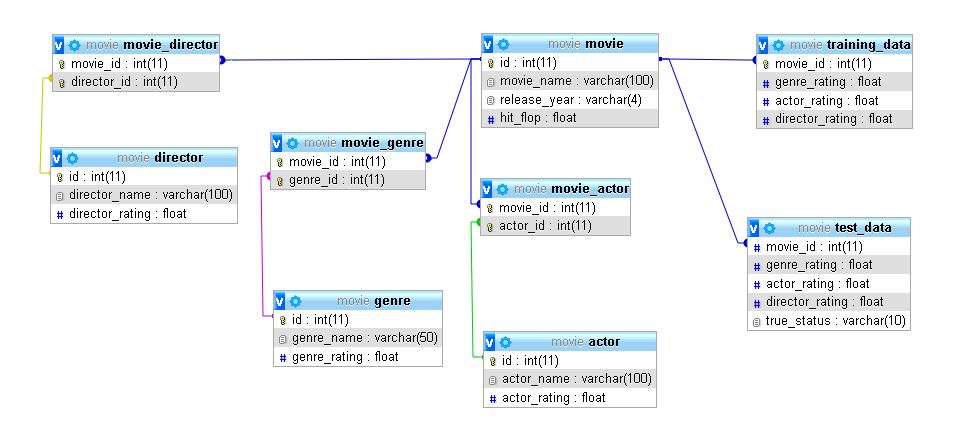


**Figure 3‑5: Level 1 DFD for process 3.0 of level 0 DFD**

# **SYSTEM DESIGN**

System design is basically a process of defining the components, modules, interfaces and data for a system in order to satisfy specified requirements. It can also be defined as a process of creating or altering systems along with the processes, practices, models and methodologies that can be used to develop them. The main objective of the detailed system design is to prepare a blueprint of a system that meets the goals of the conceptual system design requirements. The system designs used for building this project include database schema, input output design, class diagram, sequence diagram and activity diagram.

## **4.1. Database Schema**

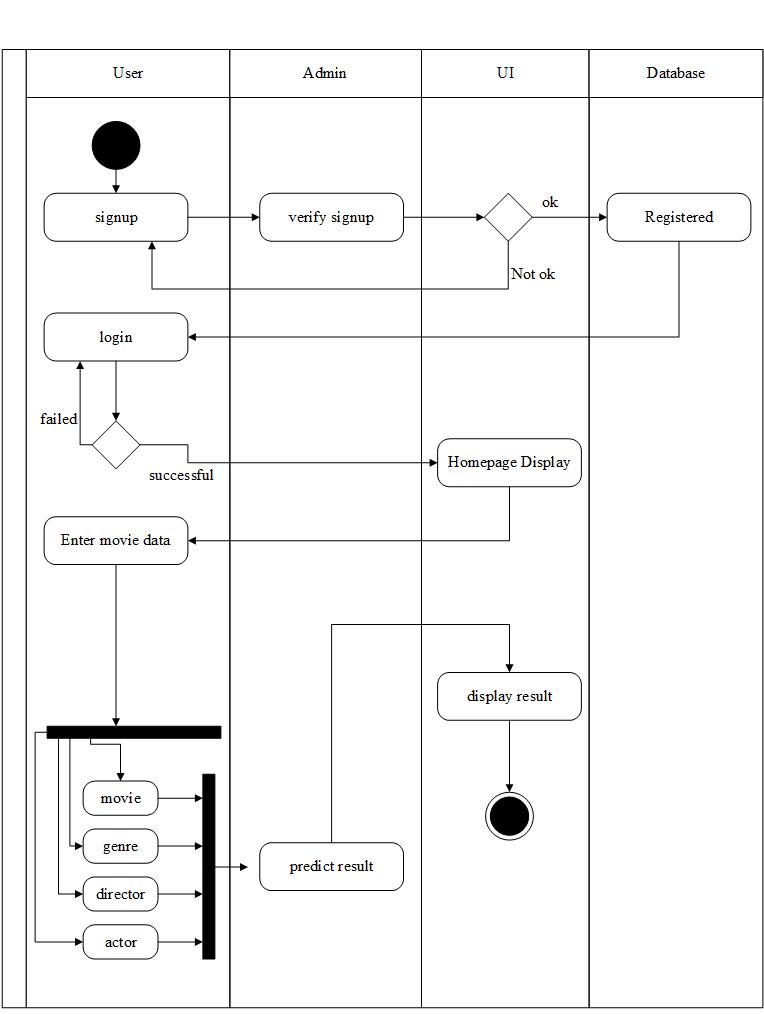
The following database schema of our application is the structure of the database used in the system in a formal language.

**Figure 4‑1: Database Schema for Movie Success Prediction System**

The above database schema of the purposed system describes the actual structure of the database. It clearly shows the implementation of nine tables namely movie, actor, movie\_actor, director, movie\_director, genre, movie\_genre, training\_data and test\_data with respective attributes.

## **4.2. Activity Diagram**

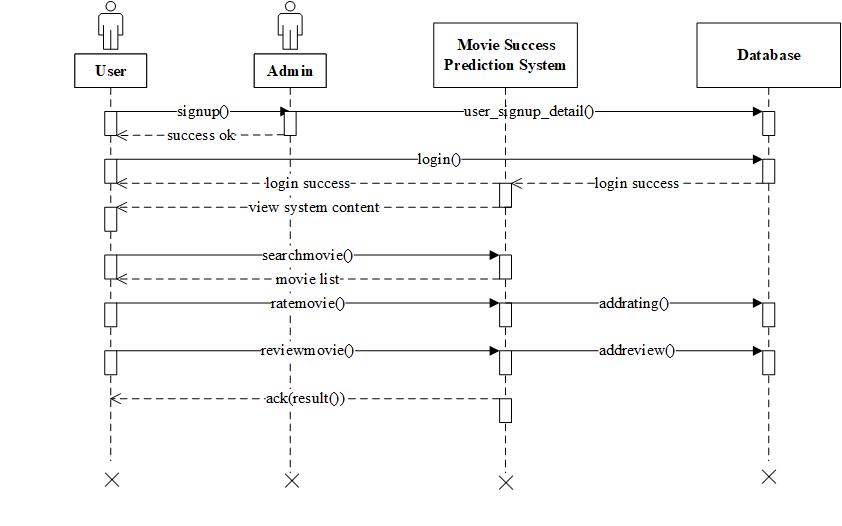
Activity diagram is important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another. The activity diagram of our system is shown as below:



**Figure 4‑2: Activity Diagram for Predicting Result**

## **4.3. Sequence Diagram**

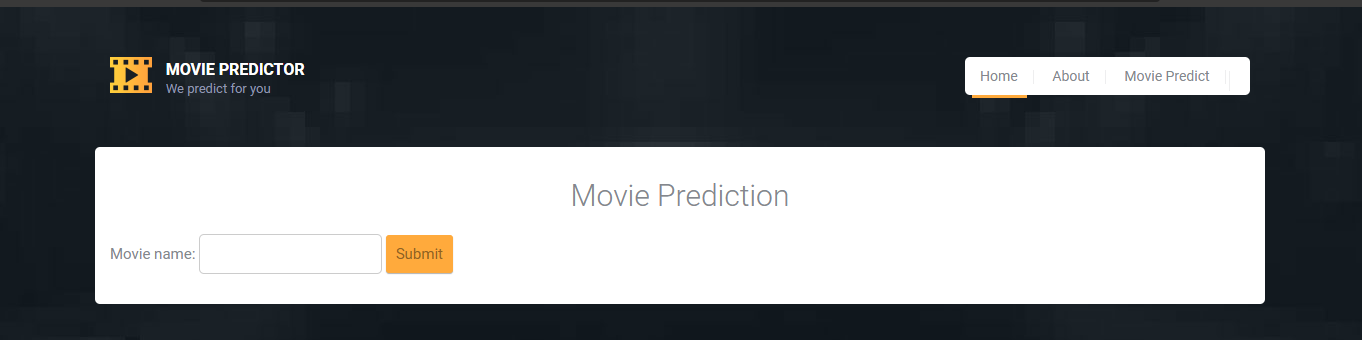
A sequence diagram is a type of interaction diagram because it describes how and in what order a group of objects works together. Sequence diagrams are sometimes known as event diagrams. The sequence diagram of our system is shown as below.



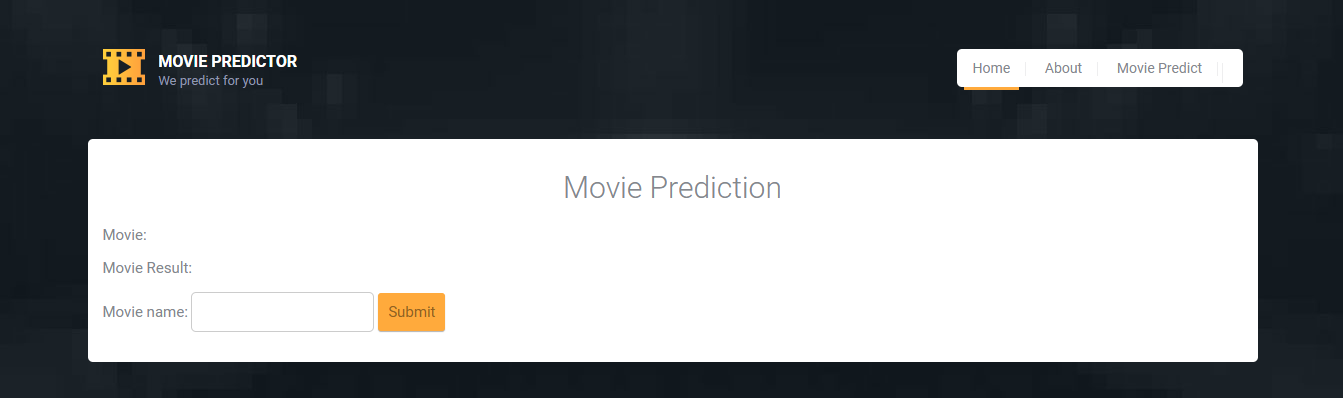
**Figure 4‑3: Sequence Diagram of User in the system**

## **4.4. Input Output Design**

Input design is the process of converting a user-oriented description of the input into a computer-based system. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. The following design shows the major input and output design of our system.



**Figure 4‑4: Input Design of User in System**

****

**Figure 4‑5: Output Design of User in System**

# **IMPLEMENTATION**

## **5.1. Tools Used**

Following are the tools used for the accomplishment of this project:

* Apache Server

Apache is the most widely used web server software, developed and maintained by Apache Software Foundation and is an open source software available for free.

* **Argon- Design System**

**An user-friendly, open source and beautiful design system based on Bootstrap 4.**

* Bootstrap

Bootstrap is a free and open-source CSS framework directed at responsive, mobile-first front-end web development which contains CSS- and JavaScript-based design templates for typography, forms, buttons, navigation and other interface components.

* CSS

CSS is a style sheet language used for describing the presentation of a document written in a markup language like HTML.

* Excel

Microsoft Excel is a spreadsheet developed by icrosoft for windows, macOS, android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called visual basic for applications.

* HTML

HTML is the standard markup language for documents designed to be displayed in a web browser.

* JavaScript

JavaScript, often abbreviated as JS, is a high-level, interpreted programming language that conforms to the ECMAScript specification.

* Microsoft Visio 2013

Microsoft Visio is a diagramming and vector graphics application and is part of the Microsoft Office family.

* MySQL

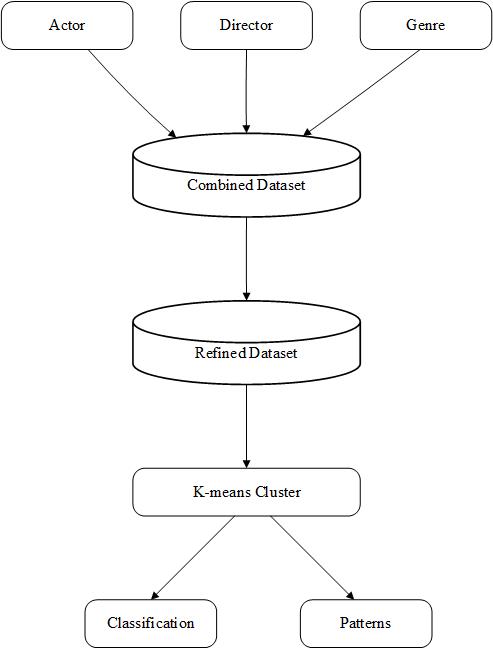
MySQL is an open-source relational database management system.

* PHP

PHP is a server side scripting language that is used to develop static websites or dynamic websites or web applications.

## **5.2. Methodology**

Figure 5.1 shows the architecture of a predictive modeling system. The data is collected for different parameters, and a database is formed. This database may contain several irrelevant information which we may not require in our predicting algorithm. Hence the database needs to be cleaned, and relevant analysis needs to be done. The database is refined and cleaned according to the requirement of the algorithm. In the proposed work, one technique is used to study dataset. It includes K-means clustering algorithm. The results of this algorithm gives us details about patterns and trends within our dataset as output. Patterns in data mining allows the users to analyze the data from different angles and dimensions. It helps us identify relationships within the data. Patterns also help in categorization and summarization of the given data. In fact, the term data mining refers to the large number of correlations and patterns found in the relational databases.



**Figure 5‑1: Architecture of Prediction Model**

The methodology has 4 major components, these are 1. Data Collection 2. Data Cleaning and Integration 3. Data Transfer 4. Data Analysis and Prediction. Cleaning the dataset and discarding the irrelevant data from the CSV dataset as well as through detailed study of the dataset, we get the attributes that can affect the prediction of success of a movie. Different algorithms are then studied so that the best suited algorithm according to our problem could be determined. The best suited algorithm should be the one which gives the most accuracy and least error. The test data when tested according to the algorithm should give as accurate results as possible.

### **5.2.1. Data Collection**

The data was collected using kaggle which contains the IMDb movie dataset of more than 1000 movies in the dataset. The data is in CSV format. The movie contained imdbid, title, releaseYear, releaseDate, genre, writers, actors, directors, sequel and hitFlop. Actor contained actorId, actorName, movieCount, ratingSum, normalizedMovieRank, googleHits, normalizedGoogleRank and normalizedRating and so as Director with directorId. Thus, some sort of cleaning, integration and preprocessing is likely to be required in order to make good use of the data for the purpose of data mining through supervised machine learning techniques. The dataset was transferred to MySQL, in form of tables.

**Table 5‑1: Bollywood Movie Detail**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **imdbId** | **title** | **releaseYear** | **releaseDate** | **genre** |
| tt0118578 | Albela | 2001 | 20-Apr-01 | Romance |
| tt0169102 | Lagaan: Once Upon a Time in India | 2001 | 8-May-02 | Adventure | Drama | Musical |
| tt0187279 | Meri Biwi Ka Jawab Nahin | 2004 | 2-Jul-04 | Action | Comedy |
| tt0222024 | Hum Tumhare Hain Sanam | 2002 | 24-May-02 | Drama | Romance |
| tt0227194 | One 2 Ka 4 | 2001 | 30-Mar-01 | Action | Comedy | Drama |

|  |  |  |
| --- | --- | --- |
| **imdbId** | **writers** | **actors** |
| tt0118578 | Honey Irani (screenplay) | Honey Irani (story) | Javed Siddiqui (dialogue) | Govinda | Aishwarya Rai Bachchan | Jackie Shroff | Namrata Shirodkar |
| tt0169102 | Ashutosh Gowariker (story) | Ashutosh Gowariker (screenplay) | Kumar Dave (screenplay) | Sanjay Dayma (screenplay) | K.P. Saxena (dialogue) | Aamir Khan | Gracy Singh | Rachel Shelley | Paul Blackthorne |
| tt0187279 | N/A | Akshay Kumar | Sridevi | Gulshan Grover | Laxmikant Berde |
| tt0222024 | K.S. Adiyaman | Arun Kumar (assistant dialogue) | Ashok Mehta (assistant dialogue) | Reema Rakesh Nath (dialogue) | Shah Rukh Khan | Madhuri Dixit | Salman Khan | Atul Agnihotri |
| tt0227194 | Sanjay Chhel | Raaj Kumar Dahima (screenplay) | Manoj Lalwani (screenplay) | Shah Rukh Khan | Juhi Chawla | Jackie Shroff | Nirmal Pandey |

|  |  |  |  |
| --- | --- | --- | --- |
| **imdbId** | **directors** | **sequel** | **hitFlop** |
| tt0118578 | Deepak Sareen | 0 | 2 |
| tt0169102 | Ashutosh Gowariker | 0 | 6 |
| tt0187279 | Pankaj Parashar | S.M. Iqbal | 0 | 1 |
| tt0222024 | K.S. Adiyaman | 0 | 4 |
| tt0227194 | Shashilal K. Nair | 0 | 1 |

**Table 5‑2: Bollywood Actor Rating**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **actorId** | **actorName** | **movieCount** | **ratingSum** | **normalizedMovieRank** | **googleHits** | **normalizedGoogleRank** | **normalizedRating** |
| 373 | Aamir Khan | 11 | 1170 | 9.448619843 | 2460000 | 7.342830181 | 10 |
| 374 | Shah Rukh Khan | 23 | 2000 | 7.550089836 | 2670000 | 7.884580135 | 9.226739883 |
| 375 | Salman Khan | 36 | 2340 | 5.402170181 | 3490000 | 10 | 9.208209991 |
| 376 | Katrina Kaif | 17 | 1640 | 8.480819702 | 2120000 | 6.46570015 | 8.948519707 |
| 377 | Deepika Padukone | 16 | 1080 | 5.64673996 | 3000000 | 8.735910416 | 8.627160072 |

**Table 5‑3: Bollywood Director Rating**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **directorId** | **directorName** | **movieCount** | **ratingSum** | **normalizedMovieRank** | **googleHits** | **normalizedGoogleRank** | **normalizedRating** |
| 1 | Rajkumar Hirani | 3 | 440 | 10 | 146000 | 2.07796 | 10 |
| 2 | Farah Khan | 4 | 440 | 7.39474 | 1060000 | 8.88317 | 8.78473 |
| 3 | Karan Johar | 5 | 430 | 5.68947 | 1050000 | 8.80872 | 7.1046 |
| 4 | Sajid Khan | 4 | 350 | 5.79605 | 905000 | 7.72912 | 6.9978 |
| 5 | Rohit Shetty | 10 | 950 | 6.32895 | 437000 | 4.24461 | 6.83761 |

### **5.2.2. Data Cleaning and Integration**

Several SOL queries were run on the relational database to clean the data in order to reduce the data and select only the relevant attributes which would help in data analysis and prediction of movie success. We used PHP to get the genre data from movie.csv and store all the required data into database. Data from different sources are mapped such as movie data and actor data to form movie and actor mapping, movie data and director data to form movie and director mapping.

**Table 5‑4: Filtered Bollywood Movie Detail**

|  |  |  |  |
| --- | --- | --- | --- |
| **imdbId** | **title** | **releaseYear** | **hitFlop** |
| tt0118578 | Albela | 2001 | 2 |
| tt0169102 | Lagaan: Once Upon a Time in India | 2001 | 6 |
| tt0187279 | Meri Biwi Ka Jawab Nahin | 2004 | 1 |
| tt0222024 | Hum Tumhare Hain Sanam | 2002 | 4 |
| tt0227194 | One 2 Ka 4 | 2001 | 1 |

**Table 5‑5: Filtered Bollywood Actor Rating**

|  |  |  |
| --- | --- | --- |
| **actorId** | **actorName** | **normalizedRating** |
| 373 | Aamir Khan | 10 |
| 374 | Shah Rukh Khan | 9.22674 |
| 375 | Salman Khan | 9.20821 |
| 376 | Katrina Kaif | 8.94852 |
| 377 | Deepika Padukone | 8.62716 |

**Table 5‑6: Filtered Bollywood Director Rating**

|  |  |  |
| --- | --- | --- |
| **directorId** | **directorName** | **normalizedRating** |
| 1 | Rajkumar Hirani | 10 |
| 2 | Farah Khan | 8.78473 |
| 3 | Karan Johar | 7.1046 |
| 4 | Sajid Khan | 6.9978 |
| 5 | Rohit Shetty | 6.83761 |
| 6 | Kabir Khan | 6.82368 |

**Table 5‑7: Filtered Bollywood Genre Rating**

|  |  |  |
| --- | --- | --- |
| **id** | **genre\_name** | **average\_rating** |
| 1 | Action | 2.33333 |
| 2 | Adventure | 2.3 |
| 3 | Animation | 1.31818 |
| 4 | Biography | 3.07692 |
| 5 | Comedy | 2.54362 |
| 6 | Crime | 2.26033 |

**Table 5‑8: Movie-Actor mapping**

|  |  |
| --- | --- |
| **movie\_name** | **actor\_name** |
| 7 1/2 Phere: More Than a Wedding | Juhi Chawla |
| 7 Khoon Maaf | Priyanka Chopra |
| 9 Eleven | Kashmira Shah |
| 99 | Soha Ali Khan |
| A Flat | Jimmy Shergill |

**Table 5‑9: Movie-Director mapping**

|  |  |
| --- | --- |
| **movie\_name** | **director\_name** |
| ...Yahaan | Shoojit Sircar |
| 16-Dec | Mani Shankar |
| 1920 | Vikram Bhatt |
| 23rd March 1931: Shaheed | Guddu Dhanoa |
| 3 Bachelors | Ajai Sinha |

**Table 5‑10: Movie-Genre mapping**

|  |  |
| --- | --- |
| **movie\_name** | **genre\_name** |
| 3 Idiots | Comedy |
| 16-Dec | Drama |
| 1920 | Mystery |
| 1920: Evil Returns | Romance |
| 2 Nights in Soul Valley | Mystery |

### **5.2.3. Data Transfer**

The relational database table data was exported to txt file for feeding to the K-means as input.

**Table 5‑11: Training dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| **movie\_id** | **genre\_rating** | **actor\_rating** | **director\_rating** |
| 2 | 2.28711 | 6.58442 | 3.74808 |
| 3 | 2.43848 | 3.62224 | 1.21331 |
| 5 | 2.34276 | 4.99397 | 2.84985 |
| 6 | 2.46038 | 5.67126 | 4.92205 |
| 7 | 2.3318 | 3.63388 | 2.90282 |

### **5.2.4. Data Analysis and Prediction**

The dataset is divided into training dataset and test dataset which contains the classes like hit, flop and average and predicting variables like actor, genre and director. K-means clustering is used to analyze the training dataset to develop models which can be used for test dataset and is also used for predicting which factors.

**Table 5‑12: Testing dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **movie\_id** | **genre\_rating** | **actor\_rating** | **director\_rating** | **true\_status** |
| 675 | 2.52718 | 5.37455 | 2.74399 | hit |
| 678 | 2.54362 | 5.5123 | 3.84718 | average |
| 681 | 2.20648 | 2.50496 | 3.04078 | flop |
| 683 | 2.24833 | 3.62047 | 1.23091 | flop |
| 685 | 2.15132 | 2.57648 | 4.72846 | flop |

**Table 5‑13: Consideration for Class**

|  |  |
| --- | --- |
| **Class** | **Rating** |
| Flop | 0-3 |
| Average | 3-6 |
| Hit | 6-10 |

The algorithms used in this system namely: K-means clustering algorithm. K-mean cluster is a used to classify the movies into three different categories. Clustering is a method, where Euclidean distance formula is used to make clusters from the data given to the algorithm. In the clustering method, clusters are formed, so that the inter cluster similarity is less, and intra cluster similarity is high. Final objective is to classify movies into three categories, we decided to form three clusters. We wanted to see, if the actor in a particular cluster, always gave the same class of movie. K means clustering aims to partition the observations into k clusters (we choose 3 here) in which observation belongs to a cluster with nearest mean serving as the prototype of the cluster. According to these algorithm, initially K centroids are chosen. Then the distance of each observation is calculated from each of those centroids using the distance formula. Below are some essential methods used in the algorithm along with its pseudocode:

**Name of Function:** InitializeMeans

**Description:** It initializes the K-means

**Inputs:** $items: tuple (actor, director, genre), $k: number of means

**Output:** $means: k centroids/means initialized with random values

**Procedure:**

$means=array();

for($i=0;$i<$k;$i++){

array\_push($means, $items[$i]);

}

return $means

**Name of Function:** EuclideanDistance

**Description:** It is used to find the Euclidean distance between two points

Distance =

**Inputs:** $x, $y: tuple (actor, director, genre)

**Output:** Distance between $x and $y

**Procedure:**

$s=0;

for($i=0;$i<sizeof($x);$i++){

$s+=pow($x[$i]-$y[$i],2);

}

return sqrt($s)

**Name of Function:** UpdateMean

**Description:** It is used to find new mean of the cluster

**Inputs:** $n: number of items in the cluster, $mean: mean of cluster to be updated, $item: the item included in the cluster

**Output:** updated mean $mean

**Procedure:**

for($i=0;$i<sizeof($mean);$i++){

$m=$mean[$i];

$m=($m\*($n-1)+$item[$i])/(float)($n);

$mean[$i]=round($m,3);

}

return $mean

**Name of function:** Classify

**Description:** It is used to classify data so that it belongs to a particular cluster

**Inputs:** $means: array of means of clusters, $item: the item to be classified

**Output:** index of the cluster in which $item lies

**Procedure:**

$minimum = PHP\_INT\_MAX;

$index=-1;

for($i=0;$i<sizeof($means);$i++){

$dis=EuclideanDistance($item, $means[$i]);

if($dis<$minimum){

$minimum=$dis;

$index=$i;

}

}

return $index

**Name of Function:** CalculateMeans

**Description:** It is used to calculate mean of clusters

**Inputs:** $k: number of means, $items: the list of items to be clustered, $maxIterations: maximum number of iterations, default value is 100000

**Output:** final means/centroids of each cluster

**Procedure:**

$means= InitializeMeans($items,$k);

$clusterSizes = array();

for($i=0;$i<sizeof($means);$i++)

$clusterSizes[$i]=0;

$belongsTo = array();

for($i=0;$i<sizeof($items);$i++)

$belongsTo[$i]=0;

for($e=0;$e<$maxIterations;$e++){

$noChange=true;

for($i=0;$i<sizeof($items);$i++){

$item=$items[$i];

$index=Classify($means,$item);

$clusterSizes[$index]+=1;

$cSize = $clusterSizes[$index];

$means[$index]=UpdateMean($cSize,$means[$index],$item);

if($index!=$belongsTo[$i]){

$noChange=false;

}

$belongsTo[$i]=$index;

}

if($noChange)

break;

}

return $means

**Name of function:** FindClusters

**Description:** It populates each cluster with the items that lie in it.

**Inputs:** $means: means of the clusters, $items: list of items to be classified

**Output:** items classified in clusters

**Procedure:**

$clusters=array();

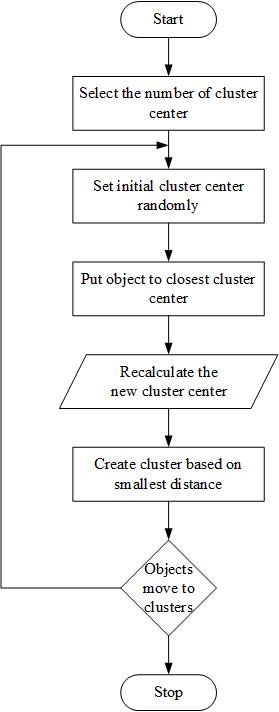
foreach ($$items as $key => $item) {

$index=Classify($means,$item);

array\_push($clusters[$index], $item);

}

return $clusters;



**Figure 5‑2: Flowchart of K-means clustering algorithm**

# **SYSTEM TESTING AND ANALYSIS**

System testing was done by giving different training and testing datasets. This test was done to evaluate whether the system was providing accurate summary or not. During the phase of the development of the system our system was tested time and again. The series of testing conducted are as follow:

## **6.1. Unit Testing**

In unit testing, we designed the entire system in modularized pattern and each module was tested. Till we get the accurate output from the individual module we worked on the same module. The input forms were tested so that they won’t accept invalid input. Finally the algorithm used was tested to predict the movie. The test case used is primary and not analogous to kaggle. Each component is tested using the test cases as follows:

**Table 6‑1: Unit Testing of Movie Prediction System**

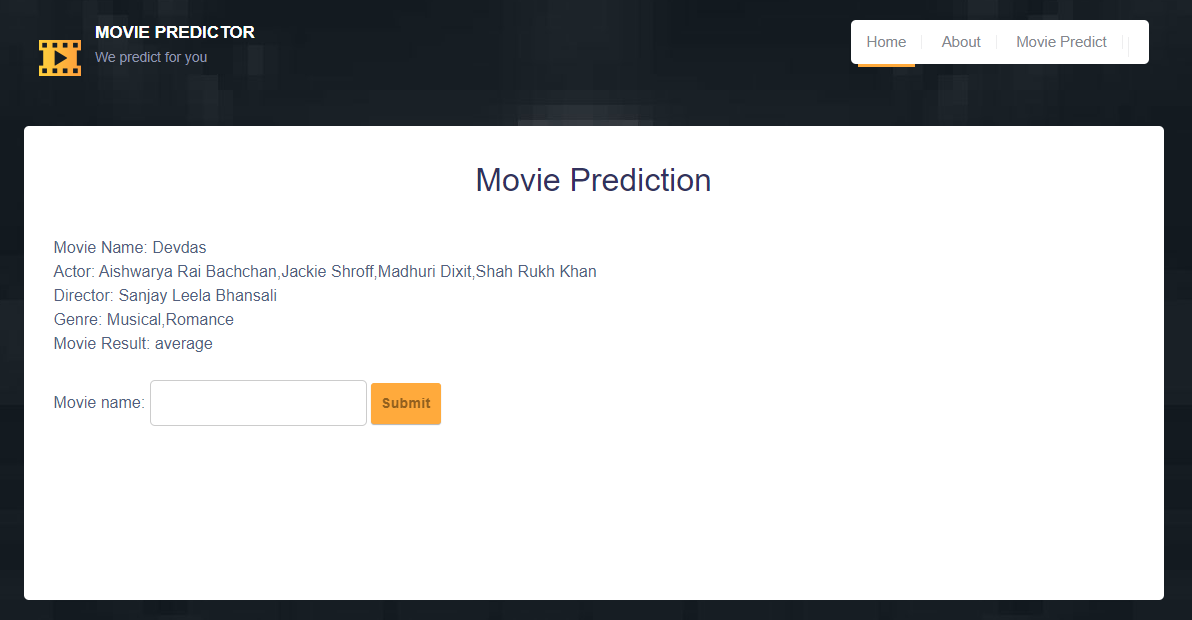
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test Case No.** | **Description** | **Test Input Data** | **Expected Result** | **Obtained Result** | **Status** |
| 1 | Euclidean Distance | Item (x) = (2.52718, 5.37455, 2.74399 )  Mean (y) = (2.318, 4.733, 3.632)  where,  x = (genre\_rating, actor\_rating, director\_rating)  y = (mean\_genre\_rating, mean\_actor\_rating, mean\_director\_rating) | Distance between x and y = 1.115304637756 | Distance between x and y = 1.115304637756 | Successful |
| 2 | Initialize Mean | List of items = (2.34747, 5.12886, 3.84718),  (2.24833, 4.72095, 4.93099),  (2.24833, 2.99504, 1.60591),  (2.27388, 4.54134, 2.74399),  (2.3318, 5.72411, 5.55705)  No. of means = 3 | Three initial means =  (2.34747, 5.12886, 3.84718),  (2.24833, 4.72095, 4.93099),  (2.24833, 2.99504, 1.60591) | Three initial means = (2.34747, 5.12886, 3.84718),  (2.24833, 4.72095, 4.93099),  (2.24833, 2.99504, 1.60591) | successful |
| 3 | Update Mean | No. of items in a cluster = 4  Mean to be updated = (4,5,6)  Item added in the cluster = (7,8,9) | Updated mean after adding item = (4.75, 5.75, 6.75) | Updated mean after adding item = (4.75, 5.75, 6.75) | successful |
| 4 | Admin Login | Email = partu@gmail.com  Password = partu | Login in successful | Login in successful | successful |
| 5 | Admin Login | Email = rahul@gmail.com Password = rahul | Login failed | Login failed | successful |

## **6.2. Integration Testing**

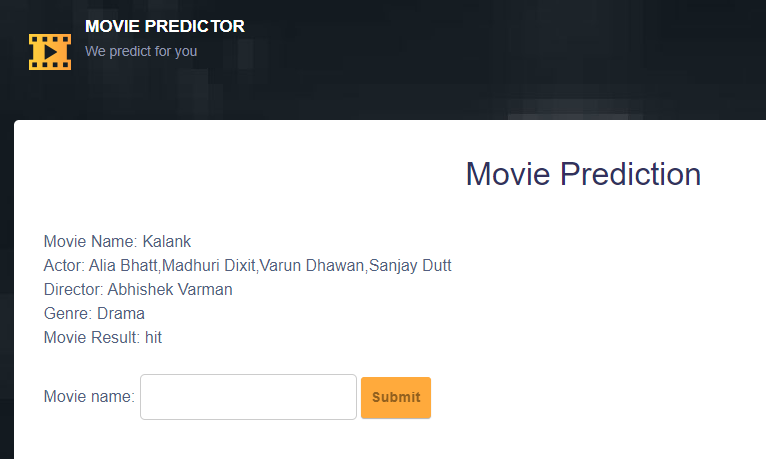
After constructing individual modules all the modules were merged and a complete system was made. Then the system was tested whether the prediction given by training dataset to testing set was correct or not

**Table 6‑2: Integration Testing of Movie Prediction System**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test Case No.** | **Description** | **Test Input Data** | **Expected Result** | **Obtained Result** | **Status** |
| 1 | Movie Prediction | Movie Name:  Devdas | Average | Average | successful |
| 2 | Movie Prediction | Movie Name: Kabir Singh | Hit | Hit | successful |
| 3 | Movie Prediction | Movie Name: Kalank | Flop | Hit | unsuccessful |
| 4 | Movie Prediction | Movie Name: 3 idiots | Hit | Hit | successful |
| 5 | Movie Prediction | Movie Name: Batla House | Hit | Average | unsucessful |



**Figure 6‑1: Test Result of Movie: Devdas**



**Figure 6‑2: Test Result of Movie: Kalank**

## **6.3. Result Analysis**

In [statistical](https://en.wikipedia.org/wiki/Statistics) analysis of [binary classification](https://en.wikipedia.org/wiki/Binary_classification), the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the [precision](https://en.wikipedia.org/wiki/Precision_(information_retrieval)) p and the [recall](https://en.wikipedia.org/wiki/Recall_(information_retrieval)) r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the [harmonic average](https://en.wikipedia.org/wiki/Harmonic_mean) of the [precision and recall](https://en.wikipedia.org/wiki/Precision_and_recall), where an F1 score reaches g its best value at 1 (perfect precision and recall) and worst at 0.

The traditional F-measure or balanced F-score (F1 score) is the [harmonic mean](https://en.wikipedia.org/wiki/Harmonic_mean#Harmonic_mean_of_two_numbers) of precision and recall:

F1 = =

**Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Precision is calculated as:

Precision = TP/TP+FP

**Recall (Sensitivity)** - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. Recall is calculated as:

Recall = TP/TP+FN

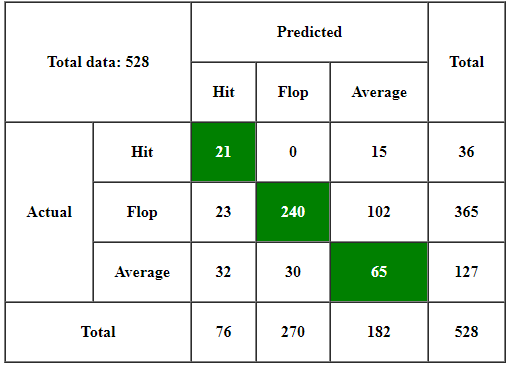
**Accuracy** - Accuracy is simply a ratio of correctly predicted observation to the total observations. One may think that, if they have high accuracy then their model is best. Yes, accuracy is a great measure but only when we have symmetric datasets where values of false positive and false negatives are almost same. Accuracy is calculated as:

Accuracy = TP+TN/TP+FP+FN+TN

Where,

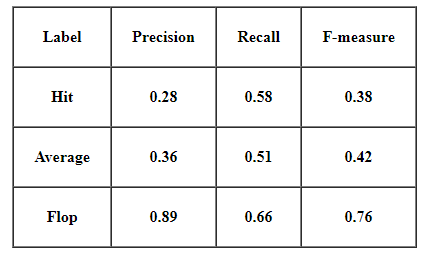
* TP: True positive
* TN: True negative
* FP: False positive
* FN: False negative

**Table 6‑3: Confusion Matrix**



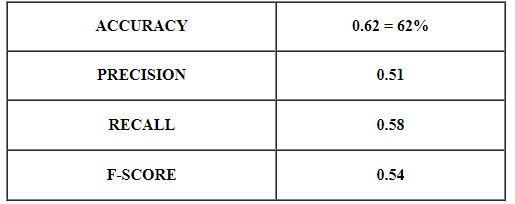
The Confusion matrix shows that out of 36 movies of category “Hit”, 21 are classified as “Hit” whereas 15 are classified as “Average”. Likewise, out of 365 flop movies, 23 are classified as “Hit”, 240 are classified as “Flop” and the remaining are classified as “Average”. Similarly, out of 127 movies of category “Average”, 32 are classified as “Hit”, 30 are classified as “Flop” and 65 are classified as “Average”.

**Table 6‑4: F-measure of different categories**

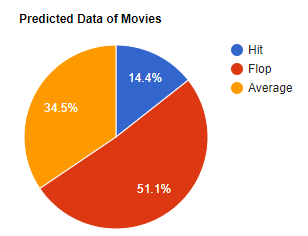
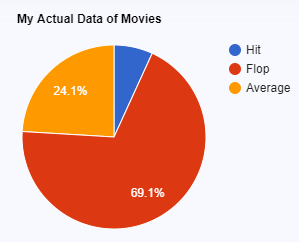


The above table shows that the precision of category Flop is the maximum, i.e. 0.89 where as the precision of category Hit is the least, i.e. 0.28. The category Average has a precision of 0.36. The recall for categories Hit, Average and Flop are 0.58, 0.51 and 0.66 respectively. The F-measure, which is harmonic mean of precision and recall, of Hit, Average and Flop are 0.38, 0.42 and 0.76 respectively. Most of the movies of the category “Flop” are correctly classified whereas most of the movies of the category “Hit” are incorrectly classified. Thus, the system predicts flop movies very well.

**Table 6‑5: Summary of result analysis**



The accuracy of the movie prediction was found to be 62% which means that out of 100 movies, 62 are correctly classified. The average F-measure of the three categories is 0.54.



**Figure 6‑3: Pie-chart of actual and predicted movie data categories**

The pie-chart above clearly shows that contrast between the actual categories of the movies and the predicted categories of the movies.

# **CONCLUSION**

## **7.1. Conclusion**

The model we used to predict movie success is quite simple, but still powerful enough to make good predictions. Compared with other proposed methods from the challenge it is by far less sophisticated, though its strengths lie in its simplicity. The major expected outcome of this system is to present users with a complete website with user interactive interface which will provide movie success/failure result to the users. The project proposed a system that predict the success of movie. In movie prediction system, training data are provided based on which result is obtained. The system uses Bollywood movie data for further prediction of Bollywood movies. The huge growth of production of movies in Bollywood and not being able to analyze the success of movie is leading to huge loss. The system can help the shareholder of movie to predict the success of their movie before investing on it. We have used K-mean algorithm for clustering of data. User can add movie detail and get the result as all work is done by data mining algorithm. Furthermore, it depends on producer or the team to whether invest on the movie or not.

# **REFERENCES**

[1] Atta Badii, Ivo Keller, Mathieu Einig, Tobias Senst, Thomas Sikora, Volker Eiselein

2013 ‘Prediction of movies box office performance using social media’.

[2] A. Kuhn, T. Senst, I. Keller, T. Sikora, and H. Theisel 2012 Comparison of Four Text

Classifiers on Movie reviews pp. 387-39.

[3] Verikas et al 2011 CART approaches to mining incomplete data IEEE Transaction on

Pattern Analysis and Machine Intelligence 34 10 pp 2064-2070.

[4] A data mining approach to analysis and prediction of movie ratings(2004) M. Saraee,

S. White & J. Eccleston

[5] Chandan Singh and Dipreet K Reddy 2014 Sentiment analysis of Indian movie review

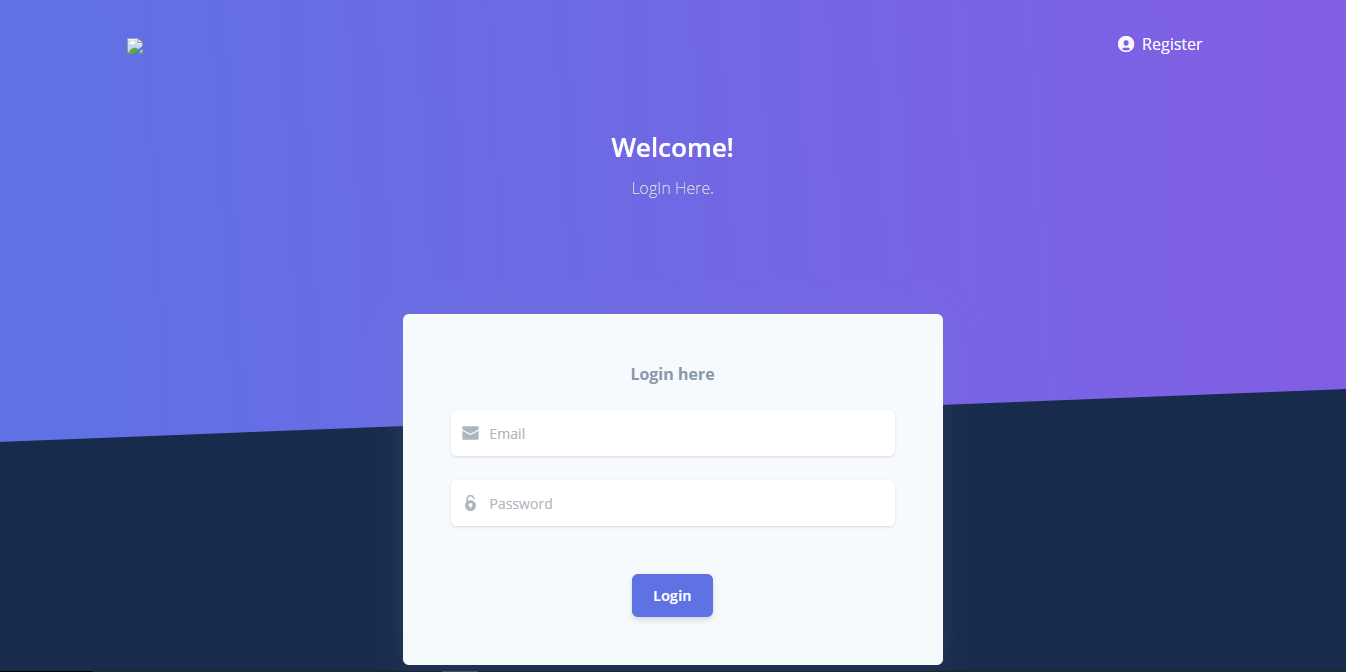
with various feature selection techniques Survey Paper,Journal of Big Data.

[6] B.Abidi,N.Aragam,Y.Yi,andM.Abidi 2008 Feature Level Sentiment Analysis on

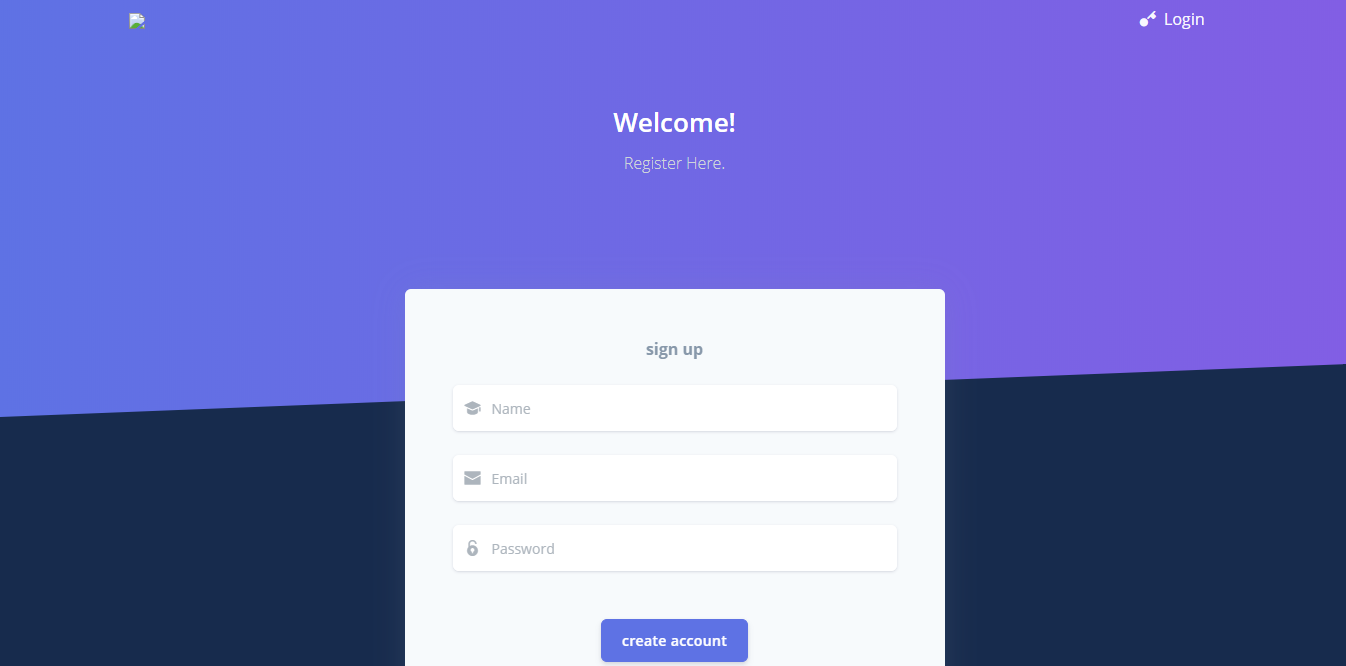
Movie Reviews ACM Computing Surveys, 41, 1 pp. 1–36

# **APPENDIX**

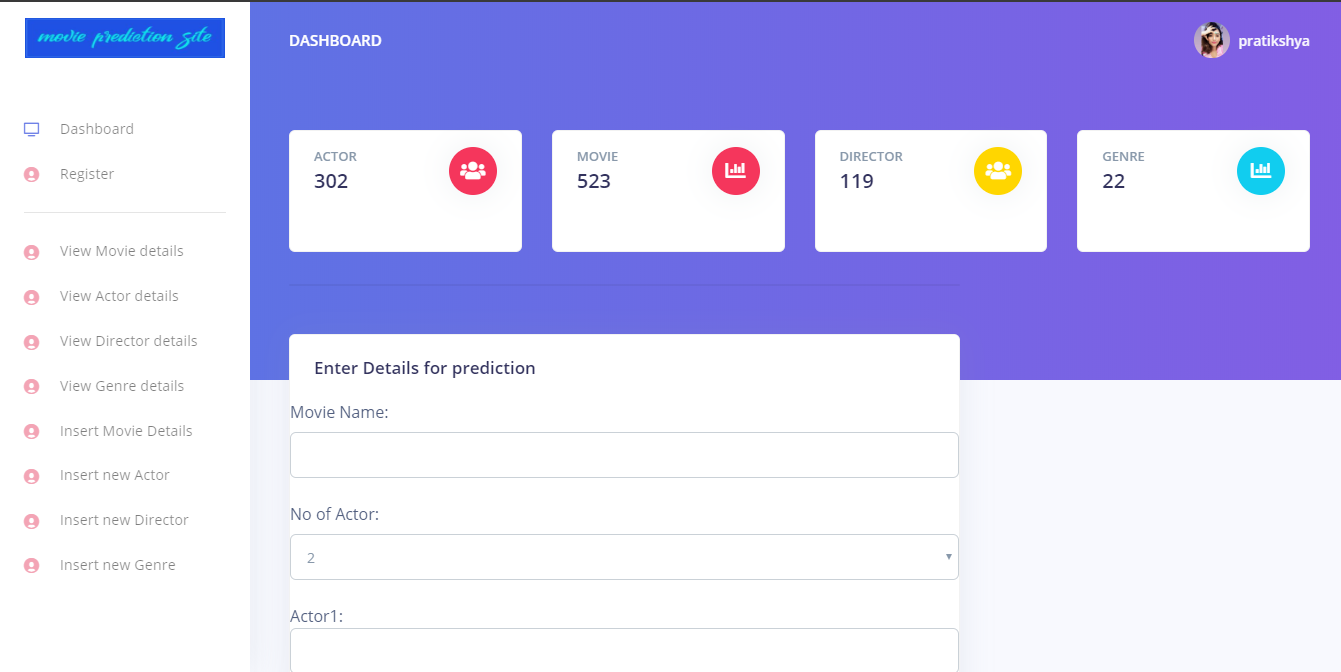
**User Interface for Admin Panel**

****

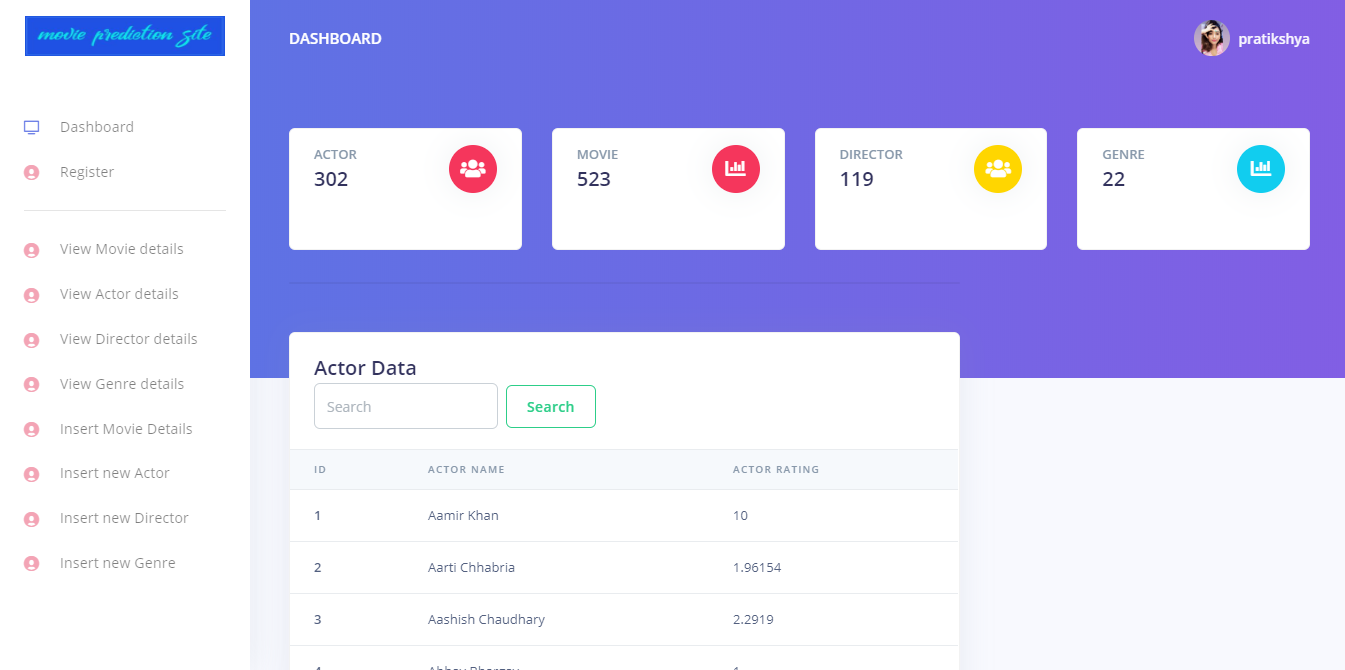
**Figure A: Login page**

****

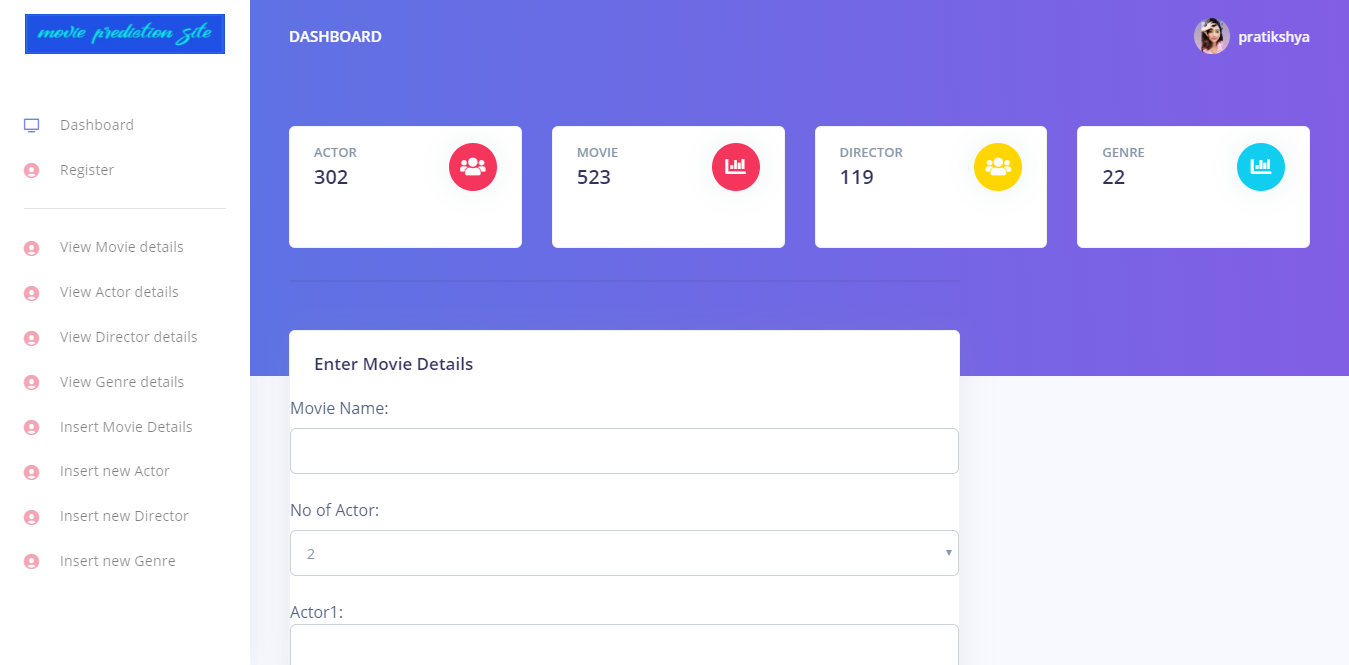
**Figure B: Registration page**



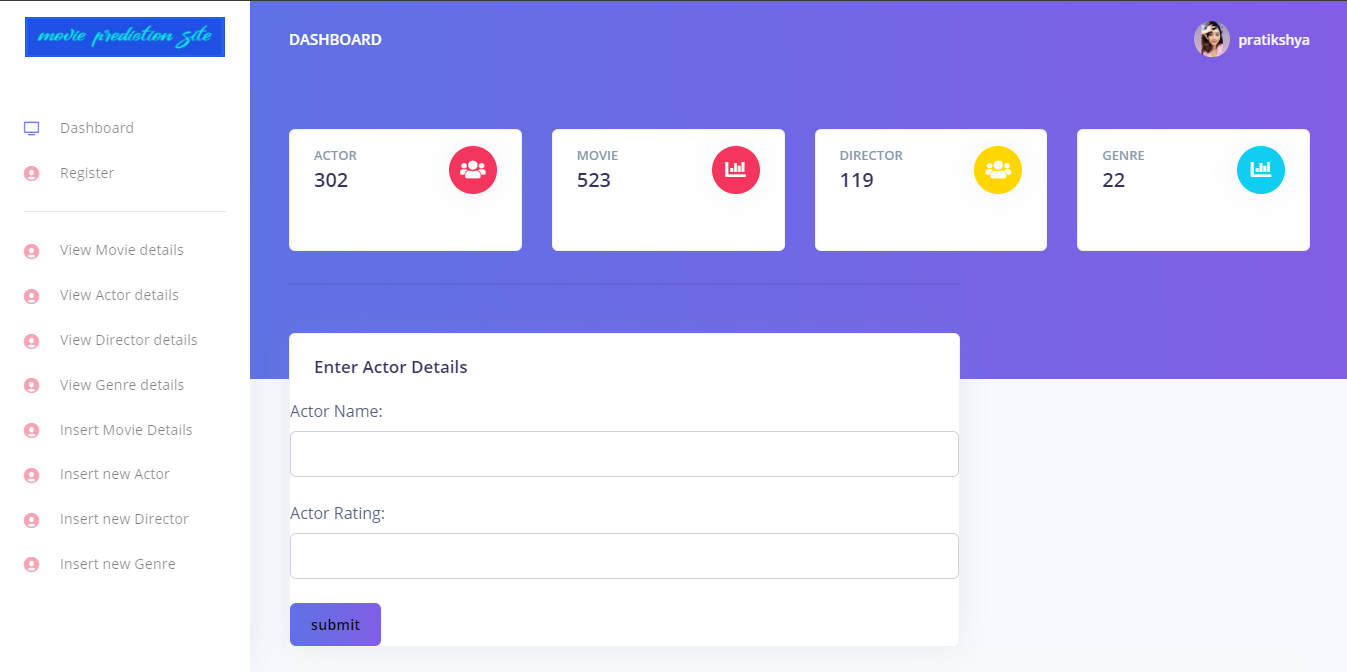
**Figure C: Prediction by Admin**



**Figure D: Search Actors details**

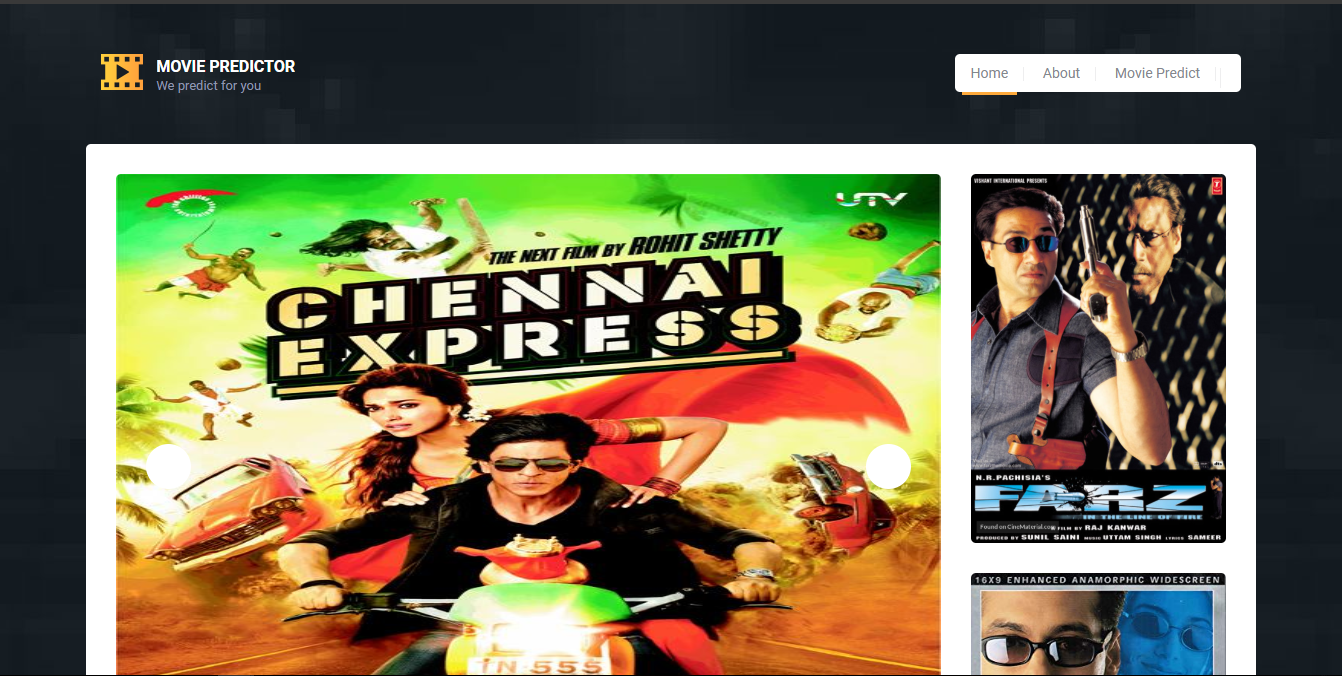


**Figure E: Movie Entry**

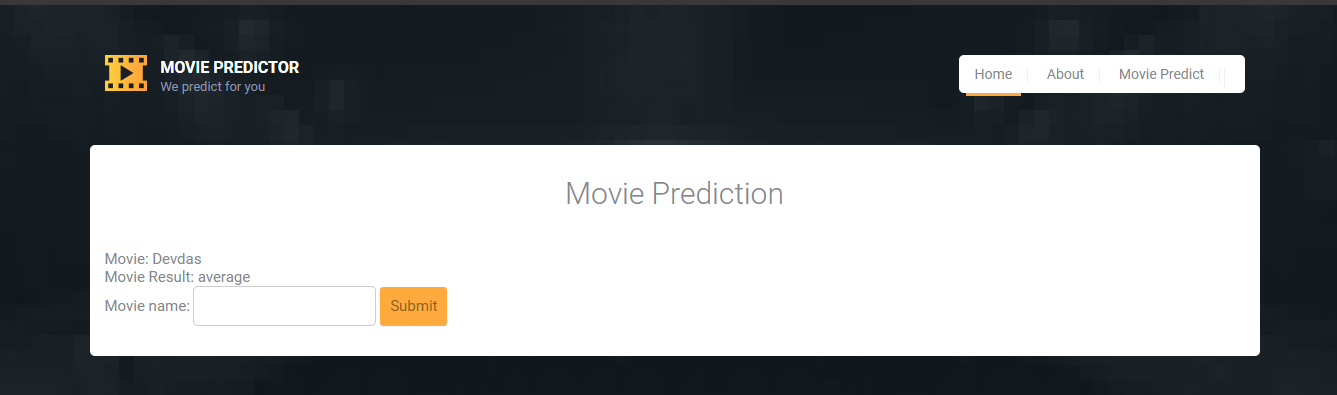


**Figure F: Actor Entry**

**User Interface for User Panel**



**Figure A: Home page of User interface**



**Figure B: Prediction page of User interface**