ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1(1), 2026

Peer Reviewed, Referred & Indexed Journal

Research Paper

Predictive Modelling of Movie Performance Using Supervised Learning and IMDb-Based Feature Analysis

Asif Ahmad[1], *Afeera Anam*[2], Nashra Kareem[3], *Saquib Ahmed*[4], Abubakar Haneef[5]

[1]Assistant Professor, Department of CSE (Data Science), Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.
[2,3,4,5]UG Students, Department of CSE (Data Science), Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.

Abstract— The movie industry requires substantial financial resources, making it crucial to anticipate a film's commercial success. This study investigates a machine learning-based method to categorize movies as HIT, AVERAGE, or FLOP using their IMDB ratings. The analysis employs three machine learning algorithms—Naïve Bayes, Logistic Regression, and Support Vector Machine (SVM)—on a real-world dataset obtained from Kaggle, which includes 5043 records with 23 attributes. The dataset is preprocessed by handling missing values, encoding categorical variables, and splitting it into training and testing sets. Experimental results indicate that Logistic Regression achieves the highest accuracy of 99%, followed by SVM with 79%, and Naïve Bayes with 35%. Furthermore, a user-friendly graphical interface has been developed, enabling users to upload datasets, train models, predict the success of new movies, and compare algorithm performance. The findings highlight Logistic Regression as a highly reliable approach for forecasting movie performance, providing a valuable decision-support tool for producers and investors.

Keywords— Movie success prediction, Naïve Bayes, Logistic Regression, Support Vector Machine, IMDB score, machine learning, dataset pre-processing, classification, predictive modelling, entertainment analytics.

I. INTRODUCTION

The entertainment industry is among the most profitable and fast-growing sectors globally, generating billions of dollars each year through film production, distribution, and related services [1]. With the rising number of films released annually, the financial risks associated with movie production are substantial. Filmmakers, producers, and investors face the challenge of committing large budgets and resources without certainty about a project's commercial performance. As a result, accurately predicting a movie's potential success has become a critical concern for industry stakeholders [2].

Movies are a distinct form of entertainment that blend storytelling, visual effects, music, and performances to captivate audiences. Their commercial success depends on multiple factors, including the quality of the storyline, cast popularity, production budget, marketing strategies, genre, release timing, and audience preferences [4]. While some films achieve blockbuster status and generate significant profits, others fail to cover production costs and are considered flops. This unpredictability makes investment in movies highly risky and emphasizes the need for data-driven prediction models that provide insights before a movie is released [5].

With the growth of digital platforms and extensive online movie databases, large volumes of data about films—such as ratings, reviews, box office earnings, and viewer demographics—are now accessible [7]. These datasets offer opportunities to apply machine learning techniques to identify patterns and correlations influencing a movie's success. Leveraging this data, predictive models can classify movies into categories such as HIT, AVG, or FLOP, assisting producers, investors, and marketers in making informed decisions [8].

The proposed movie success prediction approach employs machine learning algorithms to forecast outcomes using historical movie data and specific attributes [10]. The study uses Naïve Bayes, Logistic Regression, and Support Vector Machine (SVM) to assess performance and determine the most effective model. Logistic Regression is known for high accuracy in predicting categorical outcomes, SVM is effective at distinguishing classes in high-dimensional data, and Naïve Bayes provides a probabilistic approach assuming feature independence [11]. Comparing multiple algorithms ensures a comprehensive evaluation and selection of the best prediction method.

The dataset for this study, sourced from Kaggle, includes over 5000 movies with various features such as IMDB scores, genre, production details, and audience ratings [13]. IMDB scores are particularly useful in defining success categories, where a score of 1–3 indicates a flop, 4–6 indicates an average performance, and scores above 6 classify a movie as a hit [5]. This categorization simplifies the prediction process and provides actionable insights for stakeholders. Beyond financial planning, predictive models can guide content development by identifying elements that resonate with audiences, improve marketing strategies, and assist streaming platforms in recommending content [9]. Accurate pre-release predictions reduce financial risk and enable efficient resource allocation. Additionally, machine learning models can continuously adapt to new data, improving prediction accuracy and maintaining relevance in the evolving entertainment landscape [11].

Movie success prediction models are therefore vital tools in modern filmmaking, offering data-driven guidance for decision-making. By analyzing historical trends, audience preferences, and film-specific features, these models provide early indications of a movie's potential performance. The integration of algorithms like Logistic Regression, SVM, and Naïve Bayes ensures robust, reliable predictions [12]. Ultimately, such models empower producers, investors, and marketers to make well-informed decisions, enhancing both efficiency and profitability in movie production [13].

II. RELATED WORK

Quader et al. (2017) [1] emphasized the importance of understanding how society reacts to newly released products, particularly in terms of adoption and popularity, as a focus area for data-driven research. The film industry, with its multi-billion-dollar scale, generates massive amounts of online data. Their research introduces a machine learning–driven decision support system designed to help movie investors minimize financial risk. The system predicts potential film success by analyzing profitability trends using historical data from platforms such as IMDb, Rotten Tomatoes, Box Office Mojo, and Metacritic. Methods including Support Vector Machines, Neural Networks, and Natural Language Processing were applied to forecast box-office outcomes based on pre-release and post-release features. Results indicate that Neural Networks outperform SVM, achieving 84.1% accuracy using pre-release data and 89.27% when all features are considered, while SVM records 83.44% and 88.87% under the same conditions. Key factors influencing box-office performance include production budget, IMDb vote counts, and the number of screening locations.

Kanitkar et al. (2018) [2] discussed the transformation of Bollywood into a globally profitable market, now surpassing Hollywood in annual ticket sales. Rising production costs make predicting commercial success increasingly important. While producers can offset investment risks through satellite and music rights, distributors are still exposed if a film underperforms at the box office. Therefore, accurately forecasting a movie's success before acquiring distribution rights is critical for maintaining profitability. Their study treats movie success prediction as a supervised learning problem, employing regression techniques to estimate domestic box-office earnings and classification models to categorize films based on performance. Model effectiveness is evaluated through comparative accuracy analysis.

Raj et al. (2018) [3] highlighted that the rapid expansion of the global film industry and the significant financial investments involved make success prediction a vital task. Early predictions enable production houses to optimize marketing and promotional strategies. This study analyzed IMDb data to predict IMDb ratings using both categorical and numerical features such as director, budget, and gross revenue. The proposed approach focuses on forecasting movie success before theatrical release, minimizing reliance on subjective opinions from critics or informal sources. Multiple machine learning models were tested, with Random Forest demonstrating the highest accuracy. The study also identified factors such as user votes, critic reviews, Facebook likes, movie duration, and gross revenue as strong influencers of IMDb ratings, while genres like drama and biopic tend to perform better overall.

J. Ahmed et al. (2017) [5] proposed a comprehensive framework to estimate the likely success or failure of upcoming films by considering multiple contributing factors. Their model incorporates variables such as production budget, cast, director and producer track record, filming locations, screenwriter, release timing, competing movies, music, regional release, and target audience. Because movie production carries high financial risk, accurate forecasting is essential for stakeholders. The framework combines these factors with weighted scores reflecting their influence—for instance, lower-budget films receive smaller weights, while weekend releases are weighted higher due to increased audience attendance. Competition from major releases moderates the impact of release timing, and historical performance of actors, directors, and producers is factored into predictions. The model can also provide audiences with an indication of a movie's potential popularity before ticket purchase. Validation was conducted using simulated data, and additional considerations were proposed to enhance prediction accuracy.

III. DATASET DETAILS

The dataset utilized in this study was sourced from Kaggle and includes a detailed set of real-world movie data suitable for analysis and prediction tasks. It comprises 5,043 records and 23 features, covering both numerical and categorical information related to films. These features include details such as genre, director, cast, production budget, gross earnings, runtime, language, country, and IMDb rating. The dataset is structured with column headings in the first row, followed by individual movie entries in subsequent rows. As several attributes are non-numeric, data preprocessing is necessary before applying machine learning techniques.

Among all features, the IMDb rating plays a central role in defining the target variable. Rather than predicting the exact rating, the problem is framed as a multi-class classification task by grouping IMDb scores into categories representing movie performance. Films with ratings between 1 and 3 are labeled as FLOP, those between 4 and 6 as AVERAGE, and those above 6 as HIT. This transformation simplifies the prediction process and makes the outcomes easier to interpret for decision-makers in the film industry.

During the preprocessing phase, missing values are handled to improve data reliability, and categorical variables are converted into numerical format using label encoding methods. After cleaning and transformation, the dataset is divided into training and testing subsets, with 80% allocated for model training and the remaining 20% used for evaluation. This prepared dataset is then used to train and test Naïve Bayes, Logistic Regression, and Support Vector Machine models, enabling a balanced comparison of their effectiveness in predicting movie success categories.

IV. PROPOSED METHODOLOGY

The proposed approach focuses on predicting the success of movies using machine learning techniques applied to a structured dataset. The process follows a well-organized pipeline consisting of data collection, preprocessing, feature conversion, model development, evaluation, and prediction. This structured workflow helps in producing consistent and meaningful results while effectively handling real-world movie data.

To begin with, the dataset is obtained from Kaggle and includes a mix of numerical and categorical features such as genre, director, cast, budget, revenue, runtime, language, country, and IMDb rating. Since raw data may contain missing values and inconsistencies, an initial preprocessing step is carried out. During this phase, incomplete or null records are removed to enhance the overall quality of the dataset and avoid errors during model training.

The next step involves defining the target variable. Instead of predicting exact IMDb ratings, the problem is converted into a classification task by grouping movies into three categories: FLOP, AVERAGE, and HIT, based on specific IMDb score ranges. This conversion makes the prediction problem simpler and more useful for real-world decision-making in the film industry.

Because machine learning models require numerical input, categorical features are transformed into numeric values using label encoding. This ensures that important information from categorical data is retained while making it compatible with the algorithms. After encoding, the dataset is split into two parts: 80% for training the models and 20% for testing their performance.

The main stage of the methodology involves training three supervised learning models—Naïve Bayes, Logistic Regression, and Support Vector Machine. Each model is trained separately on the same training data to maintain fairness in comparison. Naïve Bayes acts as a probabilistic approach, Logistic Regression captures relationships between input features and output classes, and SVM works by identifying the best decision boundaries in the feature space.

In the final stage, all models are tested using the unseen test dataset, and their performance is evaluated using accuracy scores. The model with the highest accuracy is selected as the most suitable for predicting movie success. This final model is then used to classify new movie data into the defined categories. Overall, this methodology offers a reliable and scalable solution for predicting movie performance and supports informed decision-making in the entertainment sector.

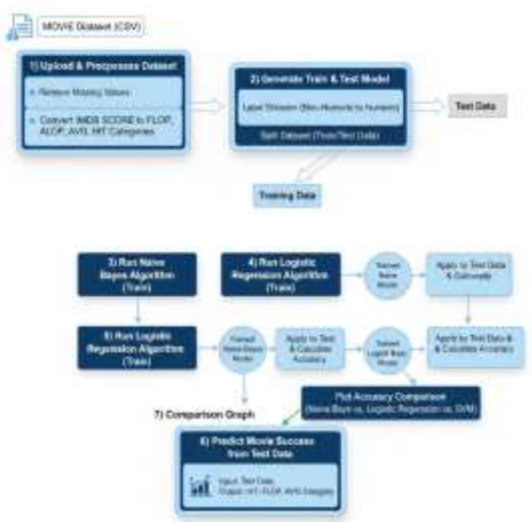


Figure 1: System Architecture of the proposed system

V.RESULT AND DISCUSSION

The performance results of the Movie Success Prediction system, implemented using Naïve Bayes, Logistic Regression, and Support Vector Machine, are demonstrated through GUI screenshots captured system execution. The experiments were conducted using a real-world dataset sourced from Kaggle, which includes 5,043 movie entries with 23 features describing various aspects such as movie details, production information, and IMDb ratings. The dataset consists of both numerical and categorical variables, making it suitable for evaluating different supervised learning models.

Figure [2] illustrates the main graphical user interface developed using Tkinter, which is launched by executing the run.bat file. This interface allows users to perform multiple operations, including uploading the dataset, preprocessing the data, generating training and testing models, executing machine learning algorithms, predicting movie outcomes, and visualizing performance comparison graphs.

Figure 2: Movie success prediction system interface

Figure [3] presents the results obtained from training and testing the Naïve Bayes classifier. The model achieves an accuracy of 35%, indicating relatively low performance. This outcome is mainly due to the algorithm's assumption that all features are independent, which may not hold true for this dataset.

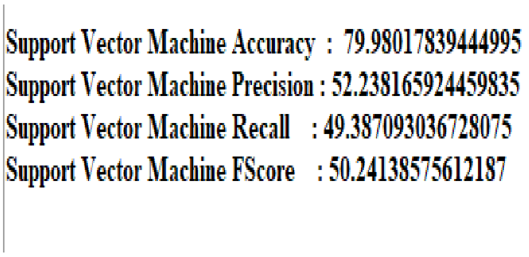
Figure 3: Naïve Bayes classification results

Figure [4] displays the performance of the Logistic Regression model during training and evaluation. The model reaches an accuracy of 99%, demonstrating its strong ability to classify movies into multiple success categories effectively.

Logistic Regression Accuracy : 99.60356788899901
Logistic Regression Precision : 99.58202716823406
Logistic Regression Recall : 99.80535279805353
Logistic Regression FScore : 99.69208646827798

Figure 4: Logistic Regression classification results

Figure [5] shows the results of the Support Vector Machine model. The SVM achieves an accuracy of 79%, performing better than Naïve Bayes but not as effectively as Logistic Regression.



Support Vector Machine Accuracy : 79.98017839444995
Support Vector Machine Precision : 52.238165924459835
Support Vector Machine Recall : 49.387093036728075
Support Vector Machine FScore : 50.24138575612187

Figure 5: Support Vector Machine classification results

Figure [6] illustrates the prediction module using unseen test data. Once the test dataset is uploaded, the trained Logistic Regression model classifies each movie as HIT, AVERAGE, or FLOP, and the predicted result is displayed along with the corresponding input details.

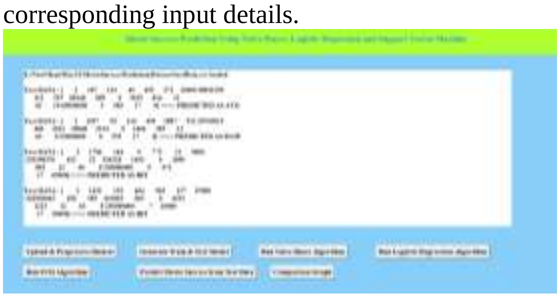


Figure 6: Movie success prediction using test data

Comparative Performance Summary

Table [1] provides a comparison of accuracy across all the machine learning models used in the system. The comparison clearly shows that Logistic Regression outperforms the other algorithms in terms of prediction accuracy.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naïve Bayes	35.97	41.45	52.73	29.66
Logistic Regression	99.60	99.58	99.80	99.69
SVM	79.98	52.23	49.38	50.24

Table 1: Model Performance Metrics

DISCUSSION

The experimental findings indicate that machine learning methods can be successfully applied to predict movie performance using historical data and IMDb-related features. Among the models tested, Naïve Bayes produces lower accuracy, mainly because its assumption of feature independence does not align well with the complex relationships present in movie data. The Support Vector Machine model delivers moderate results, as it can manage high-dimensional data and nonlinear patterns, though its effectiveness depends on parameter tuning and the nature of the dataset. Logistic Regression outperforms the other models, achieving the highest accuracy and demonstrating its effectiveness in handling classification problems where feature relationships are relatively linear.

In addition, data preprocessing steps such as removing missing values, converting categorical variables into numerical form, and appropriately splitting the dataset into training and testing sets play a crucial role in enhancing overall model performance. These results suggest that machine learning-based approaches, especially Logistic Regression, can be used as dependable tools for predicting movie success. Such models can assist industry stakeholders in making better investment decisions and planning strategies before a film's release.

VI.CONCLUSION

The movie success prediction system developed in this study highlights how machine learning techniques can be effectively used to assist decision-making in the film industry. By utilizing historical movie data along with important features such as IMDb ratings, the model is able to classify films into HIT, AVERAGE, or FLOP categories. The experimental analysis shows that Logistic Regression performs better than both Naïve Bayes and Support Vector Machine in terms of accuracy, making it the most effective approach for this dataset. The study also emphasizes the importance of proper data preprocessing and feature transformation in achieving better prediction results. Overall, this approach offers a practical, low-cost solution for producers, distributors, and investors to evaluate a movie's potential success beforehand, helping to reduce financial risks and improve planning in production and marketing strategies.

REFERENCES

1. N. Quader, M. O. Gani, D. Chaki and M. H. Ali, "A machine learning approach to predict movie box-office success," 2017

2. A. Kanitkar, “Bollywood Movie Success Prediction using Machine Learning Algorithms,” 2018
3. R. Dhir and A. Raj, “Movie Success Prediction using Machine Learning Algorithms and their Comparison,” 2018
4. J. Ahmad, P. Duraisamy, A. Yousef and B. Buckles, “Movie success prediction using data mining,” 2017
5. N. Darapaneni, “Movie Success Prediction Using ML,” 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication.
6. Pratik N. Kalamkar, Yogesh Kumar Sharma, "A Review of Research Trends in Movie Automation", 2025.
7. Irum Sindhu., Faryal Shamsi, "Prediction of IMDB Movie Score & Movie Success By Using The Facebook", 2023.
8. D. Menaga and Akshaya. L “A Method for predicting movie box-office using machine learning” Computer Science, Business, 2023.
9. R. A. Gandasari, M. Wildan, N. Anisa “Predicting Over the Top Services Movies and Shows Success Using Machine Learning”, 2023
10. N. Sikana, R. Purba “Movie Success Prediction Based on Feature and Trailer Comments Using Ensemble+LSTM Model”, 2024.
11. S. Abidi, Y. Xu “Popularity prediction of movies: from statistical modeling to machine learning techniques”, 2020.
12. N. Darapaneni, Ch. Bellarmine, K. Mondal. “Movie Success Prediction Using ML” computer science, 2020.
13. Zain Balfagih “Decoding Cinematic Fortunes: A Machine Learning Approach to Predicting Film Success”, computer science, 2024.
14. Babburi, S. Lightweight Distributed Provenance Framework for Edge and IoT Data Systems.
15. Gaddam, S. From Fixed Specifications to Self-Adapting Systems: A Machine Learning Perspective on Software Engineering.
16. Immadi, S. K. (2025). Optimizing ERP for Human Capital Management. Applied Research for Growth, Innovation and Sustainable Impact, 377–384. https://doi.org/10.1201/9781003684657-63
17. Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.
18. Poojari, R. Frameworks for Data Management and Lineage in Large-Scale Healthcare Data Systems.
19. Mahimalur, R. K., Vasgam, M., & Manoharan, D. Devops Lifecycle Management And Cloud Migration Assessments: A Security-Driven CICD Perspective.
20. Viswanathan, V. Generative AI for Smarter Workforce Planning and Enterprise Resource Decisions.
21. Gajula, S. (2026, March). Two Pillars of Banking Intelligence: A Comparative Analysis of AI Techniques for Fraud Prevention and Churn Mitigation. In 2026 14th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-6). IEEE.
22. Maturi, S. Y. (2023). Crowdsourced frontier: Unveiling autonomous adversarial cybercapabilities via open AI competition. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 275–284.
23. Venkata Ramana, P. (2024). AI-driven predictive analytics in ERP systems for proactive supply chain optimization. International Journal of Research in Information Technology and Computing, 8(4).
24. Srikanth Kavuri. (2025). AI-DRIVEN TEST AUTOMATION FRAMEWORKS: ENHANCING EFFICIENCY AND ACCURACY IN SOFTWARE QUALITY ASSURANCE. International Journal of Applied Mathematics, 38(10s), 699–710. https://doi.org/10.12732/ijam.v38i10s.990
25. Venkata Pavan Kumar Gummadi. (2026). Infrastructure Optimization Techniques for Enterprise Integration Platforms: A Comprehensive Analysis. Computer Fraud and Security, 37–44. https://doi.org/10.52710/cfs.875

26. Shashank, A. (2025). AI-Enhanced ETL Processes: Leveraging Artificial Intelligence for Optimized Data Integration Systems. Journal Of Multidisciplinary, 5(8), 219-225.
27. Kandula, S. T. R., Susarla, R. S., & Boyapati, P. K. (2025, July). Enhanced Cyber Security Using Global Local Artificial Neural Network Based Intrusion Detection in Big Data Environment. In 2025 IEEE 4th World Conference on Applied Intelligence and Computing (AIC) (pp. 426-431). IEEE.
28. Boyapati, P. K. Building a centralized data operations hub for healthcare enterprise integration. IJSAT-Int. J. Sci. Technol. 16 (2). https://doi.org/10.71097/IJSAT.v16.i2.3219