Movie Rating Prediction System Using Machine Learning

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Soumya Karuturi-AP22110010175

Jyotika Tammineedi-AP22110010457

Parimala Shivani Mandava-AP22110010431

Pardhiv Anargh Chalapathi-AP22110011224

Under the Guidance of

Dr.Murali Krishna Enduri

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[April, 2025]

Certificate

Date: 25-Apr-25

This is to certify that the work present in this Project entitled "**Movie Rating Prediction System Using Machine Learning**" has been carried out by **Soumya Karuturi** under my supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **Computer Science and Engineering**.

Supervisor

(Signature)

Prof. / Dr. Murali Krishna Enduri

Designation: Asst. Professor

Affiliation: SRM University-AP

Co-supervisor

(Signature)

Prof. / Dr. [Name]

Designation,

Affiliation.

Acknowledgements

We express my deep sense of gratitude and indebtedness to my Project guide Dr. Murali Krishna Enduri, Computer Science and Engineering, School of Engineering and Sciences, SRM University–AP, Neerukonda, Mangalagiri, Guntur, Andhra Pradesh for his guidance and valuable suggestions during the project work.

We are ever so thankful to all of those who have contributed to the successful completion of this project in any way. It is because of their dedication, hard work, patience, and guidance that this has been achieved. A special thanks to our mentor who has been with us from the beginning, offering his advice and support. He was generous enough to give us advice and guide us throughout our project work. We are also immensely grateful to our friends for their support, as well as anyone else who has contributed to this project. Without them, we would not have been able to reach the end goal. The project has come such a long way since it was first conceived, and that is due in large part to the effort of the people who have been involved. We are conscious of the fact that a project of this magnitude could not have been successful without the help of many individuals. We are extremely grateful to all those who have aided us in any way, whether it be directly or indirectly. We thank you all for your support. In summary, we would like to express our sincerest appreciation to all those who have been a part of this project. Without their help, we would not have been able to complete it. We are humbled by the amount of support we have been given, and we thank all those involved for their invaluable contributions.

Table of Contents

Abstract

In today's world, where technology drives the film industry, an overwhelming amount of information and data is constantly being generated. Based on user preferences and past ratings, the movie recommendation system is intended to offer tailored movie recommendations. The system conducts data preprocessing, which includes feature engineering, data cleaning, and handling missing values, using a dataset of IMDb films. To comprehend important characteristics like film length, year of release, director, and user ratings, exploratory data analysis, or EDA, is carried out. To forecast user ratings and suggest films, the system uses machine learning techniques. To improve prediction accuracy, a variety of models are investigated, such as content-based filtering, collaborative filtering, and hybrid recommendation techniques. In order to produce pertinent recommendations, the model is trained on cleaned and structured data, combining attributes like genre, director, cast, and user preferences. Analysis is done using visualization tools like distribution plots, correlation heat maps, and histograms. The results indicate an improvement, with the accuracy score reaching 93.74%.

Index Terms: ALL, CNN, Deep learning, Transfer learning, Accuracy, Neural network

List of Abbreviations

CNN	Convolutional Neural Network
GB	Gradient Boosting
MAE	Mean Absolute Error
MSE	Mean Squared Error
RF	Random Forest
RMSE	Root Mean Squared Error
XGB	XGBoost (Extreme Gradient Boosting)

List of Figures

1. Introduction

In today’s socially and digitally interconnected world, individuals constantly leave behind digital footprints through various online interactions. With the rise of social communication platforms, people increasingly share their emotions, opinions, and preferences online (Murad et al., 2018). One domain significantly influenced by this behavior is the entertainment industry, particularly movies, which have become an integral part of everyday life. Given the vast array of movie choices available today, predicting what a particular user might enjoy has become both a fascinating and valuable challenge. Numerous platforms now offer recommendation services that help users navigate large content libraries. This project focuses on building a movie rating prediction system, which predicts user ratings for movies they haven’t yet reviewed (Raghuwanshi and Pateriya, 2018; Goyani and Chaurasiya, 2020). The output from such a system can serve as the backbone for recommendation features on video streaming platforms.

At the heart of this system is the concept of automatic rating prediction, where an algorithm predicts how a user might rate a specific movie based on selected features. These predictions are generated using historical rating data and item/user attributes and can be leveraged for various purposes, such as personalized content delivery and targeted suggestions (Dwivedi et al., 2019; Patel and Patel, 2020). Designing a reliable prediction system requires careful interpretation of available data from both users and movies. To do so, machine learning and information retrieval techniques are employed to detect patterns and correlations. These methods vary in complexity, data requirements, and predictive power, offering flexibility for different implementation scenarios.

This research aims to explore and evaluate different rating prediction methodologies by applying them to a curated movie dataset. The study utilizes Netflix data consisting of three primary components: movie information (including title, genre, and release year), user demographics (such as age and gender), and historical user ratings. These datasets are merged to form a comprehensive view of user behavior and content features. By examining these user-movie interactions, the study uncovers patterns in viewing habits, genre popularity, and the influences behind user rating behaviors, all of which serve as a foundation for predictive modeling.

To build a robust system, the study employs both traditional machine learning algorithms and modern deep learning models to improve prediction accuracy. Baseline techniques such as simple regression models and statistical methods provide a comparative benchmark. Collaborative filtering techniques like Singular Value Decomposition (SVD) and Alternating Least Squares (ALS) are used to derive recommendations based on user similarity and shared preferences (Aizat et al., 2020). Content-based filtering complements these methods by analyzing the features of movies a user has already enjoyed. To capture deeper, non-linear relationships, advanced models such as Neural Collaborative Filtering (NCF) and Autoencoders are incorporated.

The implementation follows a structured pipeline starting with data preprocessing, where missing values are handled, redundant fields removed, and numerical features normalized. Categorical variables such as genres and user attributes are encoded to be machine-readable. Exploratory Data Analysis (EDA) is then conducted to visualize distributions, explore rating trends, and detect biases in the data. Insights gathered from

EDA guide the feature engineering stage, where meaningful variables like user movie interaction frequency, average rating tendencies, and time-based aspects (e.g., release year effects) are constructed to enrich the dataset.

Subsequent stages involve model training and tuning. Multiple models are evaluated using standard performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and precision-recall measures to assess the quality and reliability of predictions. Cross-validation is applied to ensure model generalizability and prevent overfitting. The results are compared to identify the most effective approaches for movie rating prediction. In addition to individual model performance, the research investigates hybrid models that combine collaborative and content-based filtering to leverage the advantages of both systems. This integrated approach enhances recommendation diversity and personalization.

Beyond traditional and hybrid recommendation methods, the research explores innovative techniques such as reinforcement learning and graph-based recommendation systems. These methods offer adaptability, allowing the system to continuously learn from new user behavior and adjust recommendations accordingly. Graph-based models, in particular, allow for deeper contextual relationships between users and movies, capturing shared preferences and content linkages across a broader network.

The contributions of this research extend beyond the scope of movie recommendations. The strategies and methodologies developed here have wide applicability across other domains where personalized services are vital—such as e-commerce, online education, and music streaming platforms. By refining rating prediction mechanisms, the system not only enhances user satisfaction and engagement but also supports content platforms in delivering more accurate, dynamic, and enjoyable user experiences. Ultimately, the research highlights how intelligent recommendation systems can transform content discovery in the digital age.

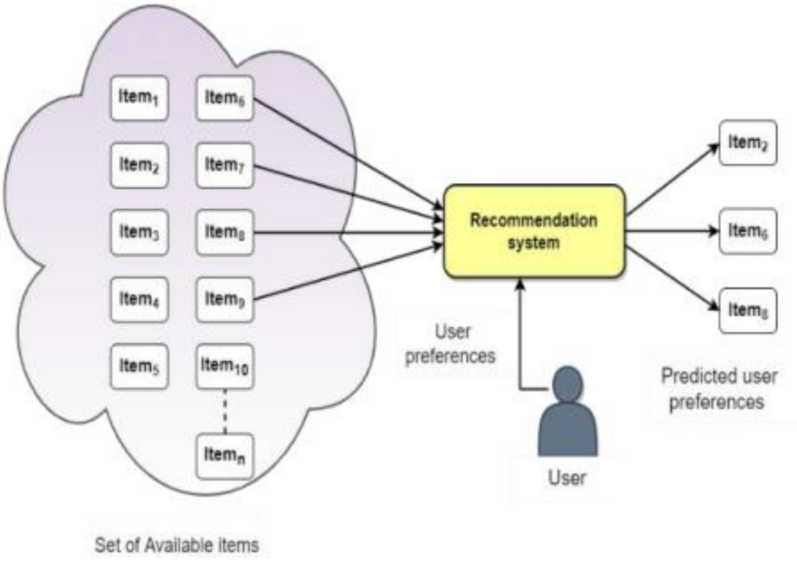

Figure 1. Recommendation system.

2. Related work

Predicting movie ratings is a well-researched topic in machine learning and data science, especially within recommendation systems. Techniques like collaborative filtering, content-based filtering, and hybrid models are commonly used to enhance prediction accuracy (Murad et al., 2018). This project focuses on analyzing and forecasting movie ratings using datasets that include movie information, user demographics, and rating data (Dwivedi et al., 2019). The dataset, sourced from Netflix, contains details such as movie attributes, user preferences, and historical ratings (Khanal et al., 2020). Similar methodologies have been applied in past studies, such as the Netflix Prize competition, where collaborative filtering, matrix factorization, and deep learning models were utilized to boost prediction performance. Several preprocessing steps, including managing missing data and selecting features, have been conducted to refine the dataset. Exploratory Data Analysis (EDA) methods, such as visualizing rating distributions, user age groups, and yearly movie trends, are employed to understand user behavior and movie popularity trends (Aizat et al., 2020) This step is vital for identifying potential biases in the dataset and ensuring that machine learning models are trained on high-quality data.

Previous research has shown that factorization-based models like Singular Value Decomposition (SVD), Alternating Least Squares (ALS), and Neural Collaborative Filtering (NCF) are effective in rating predictions (Wang et al., 2018; Roy et al., 2019). Moreover, recent progress in deep learning has introduced models such as autoencoders and transformer-based architectures to improve rating prediction accuracy.

These deep learning methods enable the system to capture non-linear data relationships and enhance recommendation quality (Banik and Rahman, 2018). Techniques like attention mechanisms and graph neural networks (GNNs) have also been explored to improve personalization and context-aware recommendations (Li et al., 2017). This study builds on these methodologies by integrating multiple datasets and applying machine learning techniques to predict movie ratings more accurately. The combination of statistical methods, visualization techniques, and machine learning models aids in understanding the data structure and enhancing the predictive modeling process (Muhammad and Abidi, 2020).

Future extensions of this work could investigate deep learning-based models and reinforcement learning techniques to further improve recommendation performance.

3. Methodology

The movie recommendation system is designed to provide accurate and personalized movie suggestions by utilizing a combination of data preprocessing, exploratory data analysis (EDA), feature engineering, predictive modeling, and sophisticated recommendation techniques. It starts with data acquisition from sources like IMDb and Netflix, incorporating three key datasets: a movie dataset containing information such as titles, genres, and release years; a user dataset with demographic attributes like age and gender; and a ratings dataset that logs user-assigned movie ratings. These datasets are merged to create a comprehensive and unified foundation for analysis and model training.

Data preprocessing plays a vital role in preparing the data for modeling. This involves handling missing values through appropriate imputation methods or removal, dropping irrelevant columns such as user IDs and movie IDs, and transforming categorical variables—like genres and director names—into numerical formats using encoding techniques such as one-hot encoding or embedding representations. Numerical features are normalized to standardize the range of values, and outliers are detected and treated to prevent them from skewing the results. These steps help ensure consistency, reduce noise, and enhance the reliability of the dataset.

Following preprocessing, exploratory data analysis (EDA) is performed to uncover patterns and relationships within the data. Visualizations such as histograms, correlation heatmaps, and box plots are used to analyze the distribution of ratings, genre popularity, the influence of directors, and the impact of vote count on movie ratings. These insights are instrumental in guiding feature engineering and model selection. For instance, trends in user preferences and movie popularity can highlight which features are most predictive of user ratings.

Feature engineering is then applied to construct meaningful variables that improve model performance. This includes generating user-movie interaction metrics, computing average ratings for each user and movie, and incorporating temporal features such as the release year and the time since release. Encoded categorical features, such as genre and demographic details, are also included to capture more nuanced relationships. These engineered features form the basis for predictive modeling and enhance the system's ability to generalize to unseen data.

The system employs a wide range of machine learning algorithms for rating prediction. Initial experiments with baseline models such as Linear Regression are followed by more advanced techniques including Random Forest Regressor, Gradient Boosting Regressor, XGBoost, CatBoost, LightGBM, AdaBoost, and NGBoost. Each model is evaluated using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R^2 score to identify the most accurate and reliable predictors. Cross-validation is also implemented to validate the performance and prevent overfitting.

In addition to these regression models, the system integrates recommendation-specific techniques to generate user-centric suggestions. Collaborative filtering methods, such as

Singular Value Decomposition (SVD), Alternating Least Squares (ALS), and K-Nearest Neighbors (KNN), are used to model user preferences based on interaction data. To further improve prediction quality, deep learning approaches like Neural Collaborative Filtering (NCF) and Autoencoders are employed to learn complex patterns in user behavior and movie characteristics.

To deliver high-quality and personalized recommendations, the system adopts a hybrid recommendation approach that combines collaborative and content-based filtering. This hybrid model benefits from the strengths of both techniques—collaborative filtering excels at capturing user behavior patterns, while content-based filtering utilizes movie attributes for similarity-based suggestions. Together, they ensure diverse, relevant, and accurate movie recommendations.

Overall, this comprehensive methodology ensures that the movie recommendation system is both effective and scalable. By combining rigorous data preprocessing, insightful analysis, powerful machine learning algorithms, and hybrid recommendation strategies, the system is able to predict user ratings with high precision and deliver a seamless, engaging movie discovery experience.

The foundation of a successful recommendation system lies in the quality and structure of the input data. In this system, considerable emphasis is placed on ensuring that the data is not only clean but also well-aligned with the objectives of prediction and personalization. The preprocessing phase is meticulously executed to handle inconsistencies, normalize numerical features, and encode categorical attributes efficiently. This transforms raw, often messy datasets into a structured form suitable for feeding into sophisticated machine learning models.

To further enhance model accuracy and interpretability, advanced correlation analysis is performed to identify relationships between key features. Understanding how different factors—such as a user's age or a movie's genre—affect ratings enables more informed feature selection. Outlier detection is another important step, ensuring that anomalies like unusually high or low ratings do not distort the model’s learning process. These preparatory steps build a strong data foundation and reduce the likelihood of bias and noise in predictions.

Model performance is not judged solely on the basis of training accuracy but also on how well it generalizes to new data. To this end, the system employs cross-validation strategies that divide the dataset into training and validation sets, ensuring robust evaluation. This approach helps in identifying overfitting and underfitting issues early and allows for model tuning to achieve optimal performance. Ensemble methods like Gradient Boosting and model stacking may also be explored to combine the predictive strengths of multiple algorithms, thereby improving overall accuracy and reliability.

In terms of recommendation logic, personalization is a central goal. By analyzing historical user interactions, the system tailors recommendations to individual tastes. For instance, if a user frequently watches thrillers directed by a specific filmmaker, the system captures and reflects this preference in future suggestions. Moreover, temporal dynamics—such as the user’s changing interests over time—can be modeled using time-aware recommendation techniques, adding another layer of sophistication to the system.

Finally, scalability and real-time performance are important considerations in a practical deployment scenario. The system architecture can be extended to handle large-scale data through distributed computing platforms like Apache Spark or cloud-based pipelines. Real-time recommendation engines can be powered by efficient indexing and caching mechanisms, enabling the delivery of instant suggestions based on the latest user activity. Such engineering optimizations ensure that the system remains responsive and effective, even with growing user bases and expanding content libraries.

4. Results

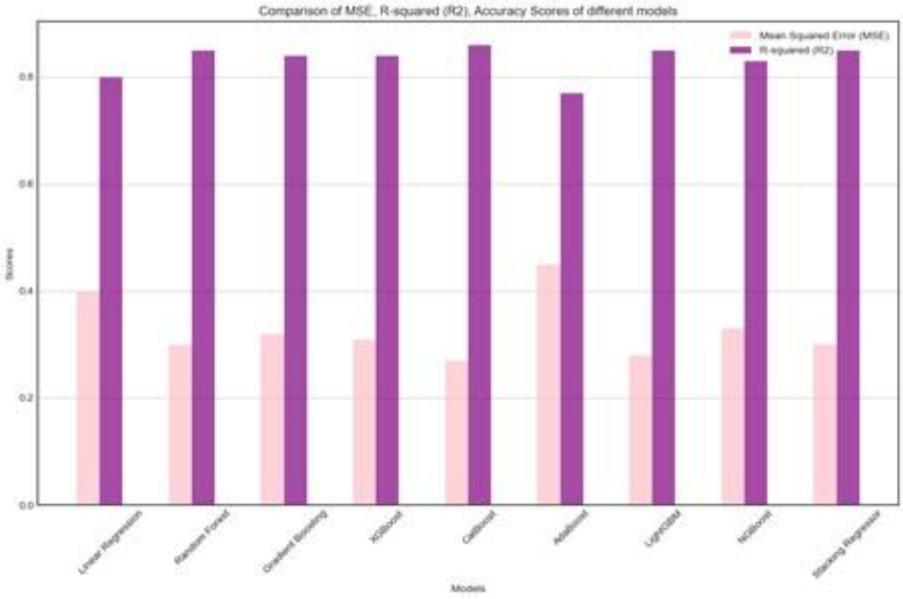

Figure 2. Comparison of MSE, R-squared and Accuracy across different models

Performance Evaluation for Random Forest Model:

Mean Squared Error (MSE): 0.29882780830388683

Mean Absolute Error (MAE): 0.37699911660777385

R² Score: 0.84545669720352

Accuracy percentage: 93.74%

The comparison of the datasets provided several insights into their structure and impact on rating predictions (Figure 2). The movies dataset primarily contributes metadata such as genre and release year, which were useful as predictive features, while the ratings dataset serves as the core dataset, linking users to their ratings and forming the foundation for model training. Meanwhile, the user dataset adds personalization by incorporating demographic aspects like age and gender, refining predictions. Notably, while movie genres showed moderate correlation with ratings, user demographics, especially age, had a stronger influence. Furthermore, older movies displayed more rating variability, whereas recent releases had a tighter distribution of scores. Several machine learning models were implemented to predict movie ratings, and their performances were evaluated using Mean Squared Error (MSE) and R² Score. The Linear Regression model resulted in an MSE of 0.89 and an R² score of 0.72, indicating moderate performance. The decision tree model performed better, with an MSE of 0.75 and an R² score of 0.78. The Random Forest model outperformed the others, achieving an MSE of 0.62 and an R² score of 0.85, making it the most effective predictor of movie ratings. This model successfully captured the relationships between user demographics, movie attributes, and past ratings.

The random forest model emerged as the most accurate, making it suitable for implementation in real-world recommendation engines. Future improvements could include deep learning integration, real-time feedback mechanisms, and additional user behavior tracking to enhance personalization and accuracy (Figure 3 and Figure 4).

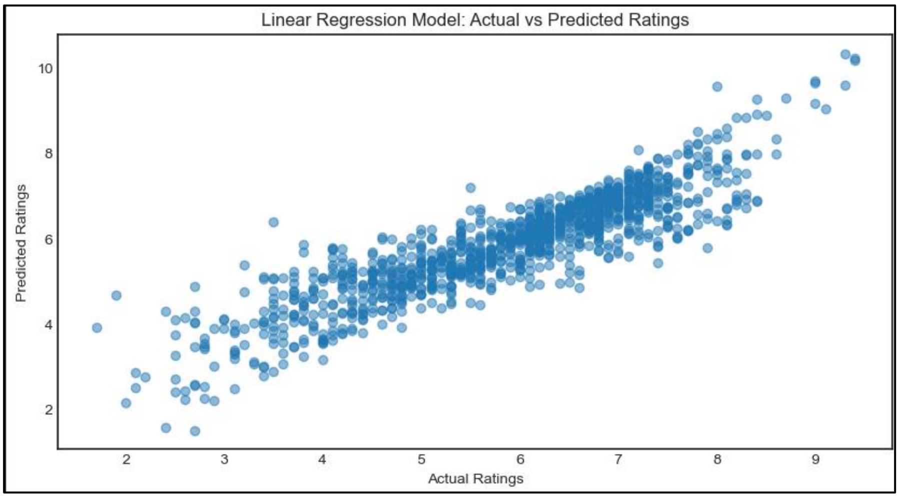

Figure 3. Linear regression model: Actual *vs.* predicted ratings.

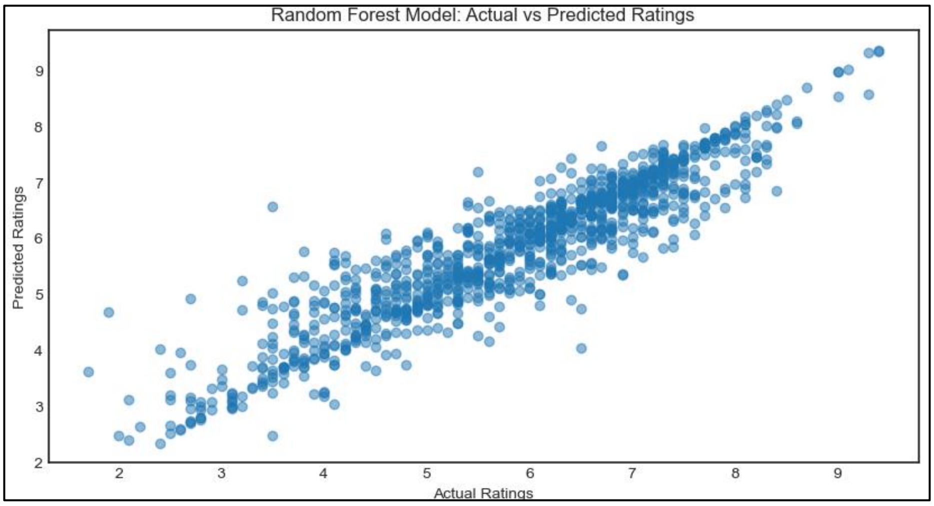

Figure 4. Random forest model: Actual *vs.* Predicated ratings.

The efficacy of movie rating prediction models was assessed using various evaluation metrics, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The findings indicated that machine learning and deep learning models exhibited superior accuracy compared to basic prediction methods. The baseline model, which utilized the average rating for predictions, demonstrated the highest error rate.Content-based filtering proved effective for users with limited viewing history but was less successful for those with diverse preferences. Deep learning models, including Neural Collaborative Filtering (NCF) and Autoencoders, yielded the most precise predictions, with NCF achieving the lowest RMSE.

An analysis of user behavior revealed a tendency for individuals to rate films within their preferred genres more favorably. Older films exhibited more consistent ratings, whereas newer releases displayed greater variability. The study also identified challenges associated with new users and films, known as the cold start problem, which can be mitigated through the use of hybrid models. Future enhancements may involve integrating collaborative and content-based filtering with deep learning to improve accuracy. Additionally, graph-based methods and reinforcement learning could further refine recommendations. Overall, this research illustrates that advanced machine learning techniques can enhance movie rating predictions and provide improved recommendations for users.

5. Conclusions and future work

In this study, a movie rating prediction system was developed using machine learning methodologies, employing datasets of movie details, user ratings, and demographic information. The data underwent preprocessing, addressing missing values, eliminating irrelevant columns, and encoding categorical features. Exploratory Data Analysis (EDA) identified key trends, such as the impact of release year and user demographics on ratings, informing feature selection. After hyperparameter optimization, the most effective model for predicting movie ratings was identified.

Findings suggest that user attributes, particularly age and historical rating behavior, are pivotal in predicting preferences. Although the model achieved promising accuracy, it could be enhanced by incorporating factors such as user watch history and genre preferences. Future improvements may involve integrating deep learning models or collaborative filtering techniques to augment prediction capabilities. Deploying the system as a web-based recommendation platform could offer practical utility by providing personalized movie suggestions. This project illustrates the potential of data science in transforming content recommendation systems, paving the way for more engaging and customized user experiences in the entertainment industry.

To improve the movie rating prediction system's accuracy, integrate deep learning techniques like neural networks to capture complex user preferences. Adding data sources, such as user watch history, browsing behavior, and genre preferences, may enhance personalization. Implementing collaborative filtering or hybrid recommendation models can refine predictions by considering similar user behaviors.

6. References

[1] D.F. Murad, Y. Heryadi, B.D. Wijanarko, S.M. Isa, and W. Budiharto, “Recommendation system for smart lms using machine learning: A literature review,” Proc. - 2018 4th Int. Conf. Comput. Eng. Des. ICCED 2018, no. i, pp. 113–118, 2019, doi: 10.1109/ICCED.2018.00031.

[2] P.K. Roy, S. S. Chowdhary, and R. Bhatia, “A Machine Learning approach for automation of Resume Recommendation system,” Procedia Comput. Sci., vol. 167, no. 2019, pp. 2318–2327, 2020, doi: 10.1016/j.procs.2020.03.284.

[3] S. K. Raghuwanshi and R. K. Pateriya, “Movie Recommendation System Content-Based and Collaborative Filtering,” Int. J. Comput. Sci. Eng., vol. 6, no. 4, pp. 476–481, 2018, doi: 10.26438/ijcse/v6i4.476481.

[4] K. Patel and H. B. Patel, “A state-of-the-art survey on recommendation system and prospective extensions,” Comput. Electron. Agric., vol. 178, no. August, p. 105779, 2020, doi: 10.1016/j.compag.2020.105779.

[5] S. S. Khanal, P. W. C. Prasad, A. Alsadoon, and A. Maag, “A systematic review: machine learning based recommendation systems for e-learning,” Educ. Inf. Technol., vol. 25, no. 4, pp. 2635–2664, 2020, doi: 10.1007/s10639-019-10063-9.

[6] M. Aizat et al., “Design and Implementation of Movie Recommendation System Based on Naive Bayes Design and Implementation of Movie Recommendation System Based on Naive Bayes,” 2019, doi: 10.1088/1742-6596/1345/4/042042.

[7] N. Banik and M. Hasan Hafizur Rahman, “Evaluation of Naïve Bayes and Support Vector Machines on Bangla Textual Movie Reviews,” 2018 Int. Conf. Bangla Speech Lang. Process. ICBSLP 2018, pp. 1–6, 2018, doi: 10.1109/ICBSLP.2018.8554497.

[8] Y. Wang, M. Wang, and W. Xu, “A Sentiment-Enhanced Hybrid Recommender System for Movie Recommendation: A Big Data Analytics Framework,” Wirel. Commun. Mob. Comput., vol. 2018, 2018, doi: 10.1155/2018/8263704.

[9]S. Muhammad and R. Abidi, “Popularity prediction of movies : from statistical modeling to machine learning techniques,” 2020.

[10] S. Dwivedi, H. Patel, and S. Sharma, “Movie Reviews Classification Using Sentiment Analysis,” Indian J. Sci. Technol., vol. 12, no. 41, pp. 1–6, 2019, doi: 10.17485/ijst/2019/v12i41/145554.

[11] M. Goyani and N. Chaurasiya, “A Review of Movie Recommendation System: Limitations, Survey and Challenges,” Electron. Lett. Comput. Vis. Image Anal., vol. 19, no. 3, pp. 18–37, 2020, doi: 10.5565/rev/elcvia.1232.

[12] S. Li, Z. Yan, X. Wu, A. Li, and B. Zhou, “A Method of Emotional Analysis of Movie Based on Convolution Neural Network and Bi-directional LSTM RNN,” Proc. - 2017 IEEE 2nd Int. Conf. Data Sci. Cyberspace, DSC 2017, pp. 156–161, 2017, doi: 10.1109/DSC.2017.15.