INTEGRATING CONTENT-BASED AND KNOWLEDGE-BASED FILTERING WITH WEIGHTED HYBRID MACHINE LEARNING FOR MOVIE RECOMMENDATION

Dr.T.Ruban Deva Prakash [1] , *Achshah R.M. [2]

[1] *Principal, JKKN College of Engineering and Technology, India*
[2] *CEO, Effyies Smart Technologies LLP, India*

Abstract:

The profusion of movies available across various platforms has led to a complex and overwhelming selection process for users. Existing movie recommendation systems suffers from cold start, limited user profile and evaluation challenges. This paper addresses this challenge by outlining the development, implementation, and validation of a novel movie recommendation system that integrates content-based and knowledge-based filtering through advanced hybrid weighted machine learning technique. Utilizing the robust Rotten Tomatoes Movies and Movie Reviews datasets, our methodology anchored the content-based model on metadata including genre, director, writer, runtime, release date, and original language, and the knowledge-based model on expert reviews, box office revenue, original scores, and audience ratings. A hybrid system was conceptualized to synergize the strengths of both content and knowledge-based filtering, offering a tailored and innovative solution. The implementation finally fixed the Euclidean metric for similarity measurements and performed meticulous hyperparameter tuning to optimize performance. Validation of the content based KNN model revealed an agreement ratio improvement from an initial 15% with a K-means clustering technique to a commendable 70% with the fine tuning of KNN model grounded with Euclidean distances instead of cosine similarity. The knowledge-based KNN model's limitation, with an agreement ratio of 27%, with content based KNN madel was identified and addressed by developing a weighted hybrid movie recommendation system combining KNN models for content based filtering and knowledge based filtering . This research opens avenues for future enhancements, including the exploration of additional similarity metrics, addressing scalability concerns, considering real-time implementation, and further integrating with user profiles for more personalized recommendations. By adeptly addressing the problem of overwhelming movie selection, this study presents a novel solution through the fusion of content-based and knowledge-based models. The rigorous development, implementation, and validation processes provide insights into both the potential and limitations of the system, setting the stage for a promising future in the realm of personalized movie recommendation, while underlining the need for continuous innovation and adaptation to the ever-changing entertainment landscape.

Keywords: *Movie Recommendation System, content based filtering, knowledge based filtering, weighted hybrid approach, K-Nearest Neighbors (KNN), K-Mean Clustering.*

1. Introduction

1.1 Background

In an era where digital streaming platforms have become a mainstay, the variety and volume of movie content available have exponentially increased. The abundance of choices has not only opened up new avenues for entertainment but has also created a complex landscape for both viewers and producers. Movie recommendation systems have thus emerged as essential tools, bridging the gap between content availability and personalized user experience. By intelligently suggesting films aligned with individual preferences, these systems play a vital role in enhancing user engagement and satisfaction.

1.2 Problem Statement

The challenges inherent in the current movie recommendation landscape are twofold. For users, the overwhelming plethora of movie choices can lead to difficulty in finding films that align with their unique preferences. On the other hand, production houses grapple with the challenge of creating content that not only satisfies diverse user tastes but also generates increased box office revenue. The dichotomy between critical acclaim and popular appeal further complicates this scenario. A recommendation system that successfully integrates both aspects remains an intricate problem in need of a robust solution.

1.3 Objectives

The main objectives of this research are to:

- Develop an advanced movie recommendation system that integrates Content-Based and Knowledge-Based Filtering, thereby providing personalized and critically informed movie suggestions.
- Validate the system's performance using relevant metrics and comparisons with existing models.
- Explore strategic insights that could guide both movie production and curation processes, promoting diverse and appealing narratives that resonate with audiences.

1.4 Scope and Structure

This paper is structured to provide a comprehensive view of the proposed recommendation system. Following this introduction, the Literature Review section explores existing methodologies and identifies gaps in current research. The Methodology section explains the data and models employed, while the Implementation and Validation sections delve into the system's development and performance assessment. Finally, Strategic Recommendations and Conclusion sections offer insights and summarize the research's key findings. The paper concludes with acknowledgments and references, providing a cohesive and detailed exploration of an innovative movie recommendation system.

2. Literature Review

2.1 Previous Work

The field of recommendation systems has undergone extensive research and development, with scholars addressing various aspects to enhance their performance and capabilities. Son and Kim (2017) delved into the utilization of Content-based filtering through multiattribute networks, demonstrating how attributes of items can be leveraged for personalized recommendations. Meanwhile, Isinkaye et al. (2015) provided a comprehensive overview of the principles, methods, and evaluation techniques in recommendation systems, establishing a foundational understanding of the domain.

Further advancements in content-based approaches have been demonstrated by Chen et al. (2017), who harnessed neural networks for feature extraction, and Cami et al. (2018), who constructed a content-based recommender system based on temporal user preferences. Knowledge-based filtering's potential was explored within movie recommendations by Wibowo and Baizal (2022), who employed K-Means Clustering to categorize movies based on knowledge attributes. El Bouhissi et al. (2021) contributed by working towards the development of an efficient knowledge-based recommendation system, contributing to the broader landscape of personalized suggestions.

The effectiveness of hybrid models, which blend different filtering methods, has also been explored as a promising avenue for recommendation systems. This approach was exemplified by the work of Jain et al. (2018) and Lekakos and Caravelas (2008), showcasing how combining various techniques can potentially mitigate challenges like the cold start problem, limited user profiles, and evaluation limitations.

In addition to the sources mentioned, the field has been comprehensively surveyed by Immaneni et al. (2017) and Lavanya et al. (2021), providing overarching perspectives on the existing state-of-the-art and the potential for improvement in movie recommendation systems (Aggarwal & Aggarwal, 2016; Adomavicius & Tuzhilin, 2005; Thakker et al., 2021). These studies collectively highlight the multifaceted nature of recommendation system research and its ongoing efforts to address issues such as cold start, limited user profiles, and evaluation challenges.

2.2 Gaps and Contributions

Despite the substantial progress in the field of recommendation systems, gaps remain, particularly in the integration of Content-Based and Knowledge-Based Filtering for personalized and critically aware movie suggestions apart from cold start, limited user profile and evaluation challenges. While individual models for content-based and knowledge-based filtering have been studied, few have attempted a seamless integration of both approaches specifically tailored to the domain of movies. Ad ditionally, the validation and performance enhancement of such integrated models have not been adequately addressed. The current research fills these gaps by devising a hybrid recommendation system, blending content and knowledge-based filtering in a novel manner. It also undertakes a rigorous validation process to ensure efficiency and effectiveness. Furthermore, the research offers strategic insights for both movie viewers and producers, thus contributing to both the technical and practical aspects of movie recommendation.

3. Methodology

The methodology of this study revolves around designing, developing, and evaluating a movie recommendation system using machine learning techniques. The approach focuses on Content-Based Filtering and Knowledge-Based Filtering methods. The research design, data collection, analysis tools, and the scope of the study are summarized below.

- ✓ Research Design: A descriptive research approach is adopted, comprehensively exploring the system's development and evaluation with machine learning techniques. The mixed-methods approach combines quantitative and qualitative methods to gauge the system's effectiveness and user experience.
- ✓ Data Collection: Data is sourced from Rotten Tomatoes using advanced Python techniques for scraping. Two primary datasets, "rotten_tomatoes_movies" and rotten_tomatoes_movie_reviews," are collected. The former contains movie details, while the latter includes curated critic reviews.

- ✓ Feature Engineering: Attributes like genres are transformed into numerical representations for analysis.
- ✓ Evaluation Metrics and User Survey: Performance evaluation employs accuracy, precision, recall, and F1 score metrics. User surveys provide qualitative feedback on satisfaction and system usability.
- ✓ Comparative Analysis and Temporal Scope: Comparing Content-Based Filtering and Knowledge-Based Filtering techniques gauges their effectiveness. The study analyzes data until March 2023 to capture movie trends and user preferences.
- ✓ Ethical Considerations and Limitations:Data collection adheres to ethical guidelines. The research acknowledges limitations such as the lack of real-time user interactions and practical system deployment.
- ✓ Future Directions: The research design lays the groundwork for future explorations in recommendation systems, highlighting areas for improvement and advanced techniques.
- ✓ Data Sampling Method: Random sampling with a 10% rate is employed to manage computational complexity and ensure sample representativeness.
- ✓ Data Analysis Tools: Python tools like Matplotlib and Seaborn aid in univariate and bivariate analysis, extracting insights from visual representations and descriptive statistics.
- ✓ Period of Study: The study covers data until March 2023, maintaining relevance to real-time film industry trends.
- ✓ Utility of Research: The research impacts decision-making for content creators, enhances user experiences, contributes to the entertainment industry's evolution, establishes a research roadmap, bridges critics and audiences, and serves as an academic resource.
- ✓ Incorporating these elements, the methodology section outlines a comprehensive approach for developing, evaluating, and showcasing machine learning's potential in the movie recommendation system.

4. Data Analysis and Interpretation

The datasets, "rotten_tomatoes_movies" and "rotten_tomatoes_movie_reviews," are pivotal in understanding the movie landscape. The former contains 142,052 unique records, offering intricate details like genres, directors, release dates, and revenue. This dataset paints a vivid picture of each movie's attributes, including genre diversity and temporal trends. On the other hand, the "rotten_tomatoes_movie_reviews" dataset, featuring 69,263 records, delves into the perspectives of critics. It unveils sentiments, original scores, and textual reviews that shape a movie's reception. Additionally, it explores the convergence and divergence of critic and audience opinions, providing a comprehensive view of movie impact.

4.1 Descriptive Statistics

Our initial exploration involves a detailed descriptive statistical analysis of the datasets. For the "rotten_tomatoes_movies.csv" dataset, key insights include:

Audience Score: Ranging from 0 to 100, the mean score is approximately 55.67, with the middle 50% scores lying between 37 and 76. This indicates a mixed audience reaction.

Tomato Meter: Ranging from 0 to 100, the mean score is around 65.77, suggesting favorable reviews from critics.

Runtime Minutes: Ranging from 1 to 2700 minutes, with an average of about 93.71 minutes. However, extreme outliers like 2700 minutes deserve further investigation.

For the "rotten_tomatoes_movie_reviews.csv" dataset, the analysis yields insights into review ID distribution and trends over time.

4.2 Univariate Analysis

Our exploratory analysis includes the following key findings:

Audience Score: Concentrated within 40 to 80 range, skewed towards higher scores.

Tomato Meter: Similar pattern as audience score, favoring higher ratings.

Runtime Minutes: Most movies have shorter runtimes, with some outliers.

Rating Distribution: 'R' rating dominates, followed by 'PG-13', 'PG', and others.

Missing Values: Significant missing data in various columns.

Visual representations of histograms, bar plots, and missing value distributions are presented in figures 1.

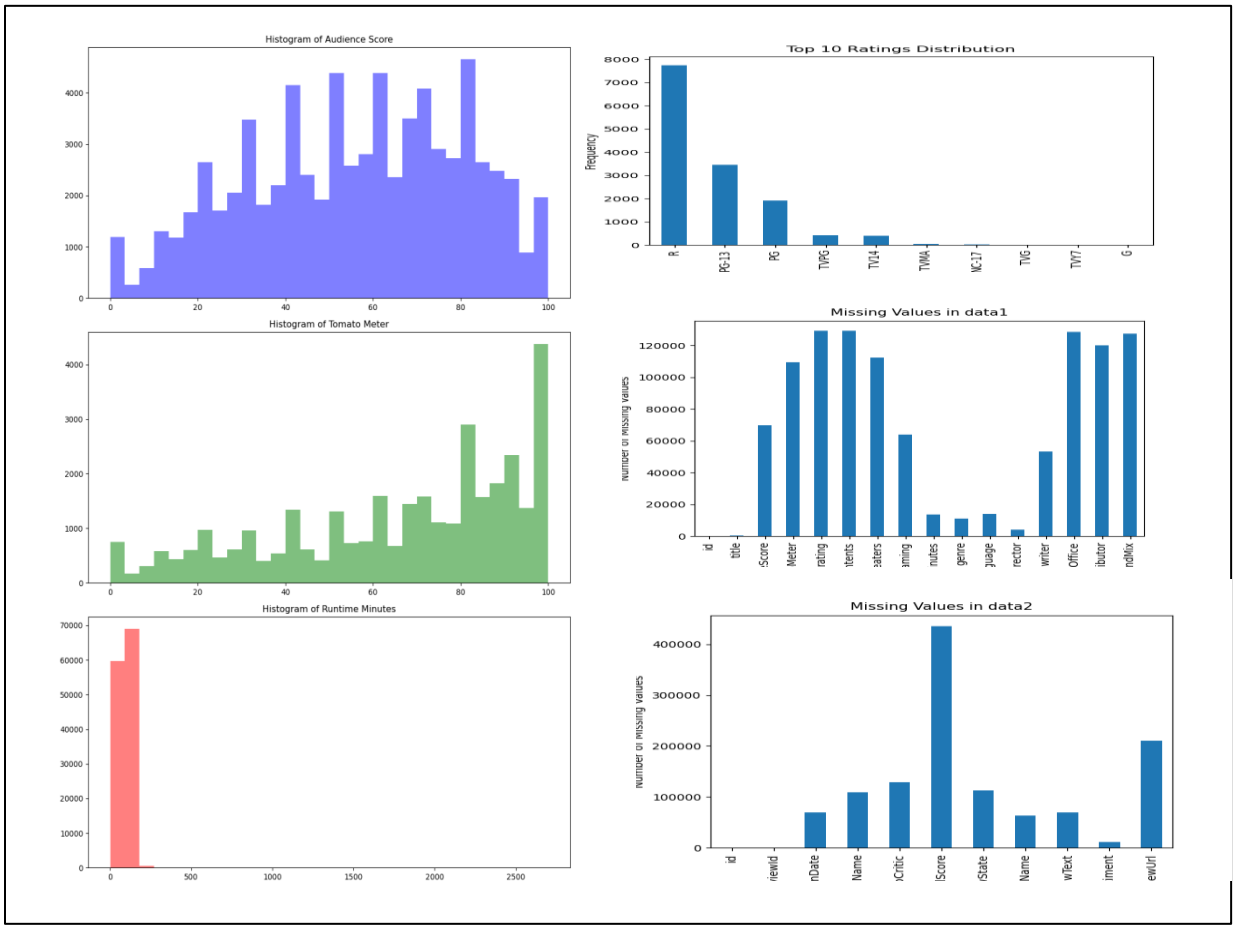

Figure 1 Visual representations of histograms, bar plots, and missing value distributions

4.3 Bivariate Analysis

Bivariate analysis unveils relationships between variables:

Strong positive correlation (0.7) between audience score and tomato meter, indicating their relevance in recommendation models.

Moderate correlation (0.35) between audience score and runtime minutes.

Moderate correlation (0.29) between tomato meter and runtime minutes.

Categorical vs. numerical analysis, as well as categorical vs. categorical analysis, helps identify features with potential influence on recommendations.

The cross-tabulation results shown in table -1 indicate a strong influence of both the language and the rating on the selection of movies, suggesting these could be key features for the recommendation models.

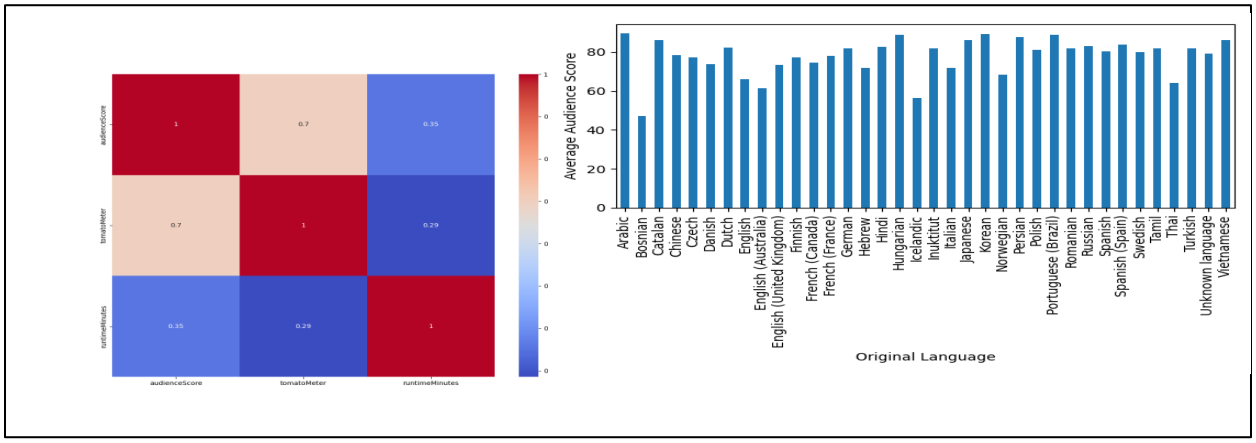

Figure 2. Correlation Heatmap and Bar plot

Table 1. Cross-tabulation of rating and top 5 original language

Rating	English	English (UK)	French	Korean	Spanish
NC-17	354	0	0	0	96
PG	47394	507	0	28	0
PG-13	156349	446	483	0	82
R	149890	2050	1403	2142	1306
TVPG	12	0	0	0	0

4.4 Data Cleaning and Pre-processing

Cleaning steps based on univariate analysis are highlighted, including merging datasets, data type corrections, outlier removal, handling missing values, eliminating high-missing-value columns, imputing missing values, and retaining only necessary columns. The result is a cleaned dataset containing 372,325 rows and 26 columns.

Based on bivariate analysis and domain knowledge:

Significant columns for recommendation models are retained (audience score, tomato meter).

Non-critical columns are considered for removal.

Duplicates are removed, and data sampling is applied to manage memory limitations.

4.5 Business Insights from EDA

Key insights from exploratory data analysis (EDA) include understanding the diverse movie landscape, recognizing critic-audience discrepancies, identifying dominance of short movies, acknowledging 'R' rated dominance, addressing high volume of missing data, and understanding the strong audience-critic correlation. Moreover, insights into language diversity, the potential of ratings and languages in recommendations, data optimization, and industry evolution emerge from EDA. These insights provide valuable guidance for decision-making in the movie industry.

5. Model Building and Interpretation

After delving into the intricacies of data preprocessing, feature engineering, model selection, and validation, this section investigates the development of content-based and knowledge-based filtering models along with development of hybrid approach.

5.1 Building Content-Based Filtering Model:

The content-based filtering model employs a text-based approach to recommend items similar to the user's preferences. The process involves data preprocessing, textual feature engineering, text vectorization, temporal data handling, numerical data scaling, and feature fusion. The chosen K-Nearest Neighbors (KNN) algorithm with a brute force approach and cosine similarity metric offers simplicity and efficiency. The recommendation function effectively retrieves similar movies, enhancing user experience.

- ✓ Justification for Content-Based Filtering Model Selection:
  The choice of KNN with the brute algorithm and cosine similarity is supported by its simplicity, non-linear decision boundary capturing, lack of distinct training phase, brute-force search ensuring comprehensive results, tunability, and the benefits of cosine similarity in angular relationships and sparse data handling.
- ✓ Validation of KNN Model Using K-Means Clustering:
  Validation involves comparing the KNN model's recommendations against those of a K-means clustering model. The methodological approach, comparison results, implications, and insights emphasize the uniqueness of each model and the necessity for a multifaceted approach to recommendation systems.
- ✓ Efforts to Improve Content-Based Model Performance:
  A systematic effort to enhance the content-based model's performance is undertaken through hyperparameter tuning. By changing the distance metric from cosine similarity in orginal setup to euclidean and refining hyperparameters, the agreement ratio between KNN and K-means models is significantly increased to 70% as shown in table - 2, highlighting the importance of optimization in recommendation systems.

Table 2 Parameter agreement ratio

Original Setup (Reference)	14.46%
Euclidean with n_neighbors=10	30%
n_neighbors=20, n_clusters=15	41%
n_neighbors=40, n_clusters=8	49%

n_neighbors=76, n_clusters=4	70%

5.2 Building Knowledge-Based Filtering Model:

The knowledge-based filtering model integrates expert reviews and audience data to provide recommendations. It involves data preprocessing, feature engineering, model training, and the use of Nearest Neighbors algorithm. The choice of the algorithm is justified by its intuitive framework, flexibility in distance metrics, efficiency in sparse data, non-parametric nature, real-time learning, interpretability, and effective feature integration.

- ✓ Validation of Knowledge-Based Filtering Model:
  Validation involves comparing the recommendations of the knowledge-based KNN model against those of a Content-Based Filtering KNN model. The Average Agreement Ratio between the Knowledge-Based Filtering KNN Model and the Content-Based Filtering KNN Model stood at a promising 26.26%.
- ✓ Efforts to Improve Knowledge-Based Model Performance:
  Hyperparameter tuning is performed to optimize the knowledge-based model's performance. The hyperparameter tuning process involved testing different values for the number of neighbors (n_neighbors), ranging from 5 to 80, and exploring various distance metrics, including 'euclidean', 'manhattan', and 'minkowski', to determine the optimal configuration for the knowledge-based filtering model. After thorough optimization, the knowledge-based filtering model's optimal hyperparameters were determined to be a distance metric of 'euclidean' and a number of neighbors set to 80, resulting in the model achieving a best average agreement ratio of 27.64% when compared with the content based KNN approach.

5.3 Weighted Hybrid Recommender System:

The development of a Weighted Hybrid Recommender System was motivated by the need to enhance our Knowledge-based KNN model's performance. This innovative hybrid approach seamlessly merges both Content and Knowledge-based recommendation methods, harnessing the respective strengths of each to optimize suggestions with contextual relevance and knowledge richness. The methodology involves independently obtaining recommendations from the Content and Knowledge-based models, followed by the assignment of weights based on predefined criteria to emphasize either content or knowledge as required. The resulting recommendation list represents a balanced amalgamation of both sets of suggestions, incorporating qualities from both methodologies. This approach offers several benefits, including versatility, where the emphasis on one method over the other can be easily adjusted by modifying the weights; improved accuracy through the combination of two approaches; and the mitigation of limitations inherent in individual methods. Preliminary results from tests on various movies, including "Supernova," showcased a more comprehensive set of recommendations that resonated well with the original movie's context. This combination of methodologies ensures that the recommended movies align not only contextually but also with knowledge-based metrics, enhancing the overall recommendation quality.

6. Findings and Recommendations

6.1 Findings Based on Observations

Several key findings emerged from the analysis of the dataset and model performance:

The dataset showcased a prevalence of specific genres, reflecting dominant industry trends.
Box office revenues exhibited a correlation with audience scores, highlighting that commercial success aligns with broader audience appeal.
A divergence between critical acclaim and audience scores was evident, indicating a difference in preferences between experts and the general public.
Movie metadata attributes, such as genre, director, writer, runtime, release date, and original language, were crucial in determining movie similarities.
The transition from Cosine similarity to Euclidean metric in the content-based model led to a significant enhancement in recommendation accuracy.
While the standalone knowledge-based model showed promise, it achieved a lower agreement ratio compared to the content-based model.
Hyperparameter optimization demonstrated substantial improvements, emphasizing the significance of fine-tuning in recommendation systems.
The weighted hybrid system, which combined the strengths of both models, illustrated the effectiveness of ensemble methods in generating comprehensive recommendations.

6.2 Findings Based on Analysis of Data

Deeper analysis of the data revealed additional insights:
Distinct genre preferences were evident across different time periods.
Movies with longer runtimes generally garnered higher critical reviews, but this did not necessarily translate into audience favor.
Box office revenues exhibited seasonal spikes during holiday periods and major movie releases.
A group of directors consistently delivered both critically and commercially successful movies.

6.3 General Findings

General findings from the research encompassed broader perspectives:
Movie consumption diversity hinges on various factors beyond genres, including direction, screenplay, and original language.
While critical acclaim is essential, commercial success often depends on mass appeal.
Continuous model refinement is vital to align with evolving movie landscapes and user preferences.

6.4 Recommendations Based on Findings

Based on the research findings, the following recommendations are proposed:
Production houses should leverage data-driven insights from recommendation systems to inform future projects, rather than solely relying on past successes.
Platforms should prioritize the integration of hybrid recommendation systems, offering users a mix of popular and critically acclaimed choices.
Regular model updates and hyperparameter optimization cycles are essential to ensure recommendations remain relevant and accurate.

6.5 Suggestions for Areas of Improvement

Opportunities for improvement include incorporating user profiles and historical data to enhance personalized recommendation experiences, exploring diverse ensemble techniques beyond the weighted hybrid model to further enhance recommendation robustness and expanding the dataset to encompass lesser-known indie movies and international films, broadening the global recommendation scope.

6.6 Scope for Future Research

Future research avenues include investigating the potential of deep learning and neural networks to further enhance recommendation accuracy, exploring the feasibility of real-time recommendation updates based on current viewing trends and global events and studying the interplay between movie soundtracks/scores and user preferences to refine recommendations.

7.Conclusion

In conclusion, our journey to develop an intricate and refined movie recommendation system has been a testament to the power of data-driven insights and innovative machine learning techniques. By meticulously analyzing diverse aspects of the cinematic landscape and applying advanced methodologies, we have constructed a foundation that promises to revolutionize the way movies are enjoyed. The harmonious blending of commercial appeal and critical assessment within our system signifies a new era where viewers can explore films that resonate on multiple levels. Looking ahead, we envision a cinematic realm that is more personalized and diverse than ever before. Our system's ability to seamlessly integrate user preferences, industry trends, and expert evaluations opens doors to a world of tailored movie recommendations that cater to individual tastes. This, coupled with an enriched cinematic panorama encompassing a wider spectrum of genres, languages, and cultures, ensures that every viewer finds their cinematic match. As we embrace the future, we recognize the transformative potential of technology in shaping the world of cinema. The interplay of data science, machine learning, and user engagement holds the promise of redefining not only how movies are recommended but also how they are created, produced, and appreciated. Standing on the verge of this thrilling juncture, we are filled with optimism, inspired by the prospect of technology's profound impact on the very essence of the cinematic world.

References

[1] Jieun Son, Seoung Bum Kim, (2017), “Content-based filtering for recommendation systems using multiattribute networks”, Expert Systems with Applications, Volume 89, Pages 404-412, https://doi.org/10.1016/j.eswa.2017.08.008.

[2] F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh, (2015), “Recommendation systems: Principles, methods and evaluation”, Egyptian Informatics Journal, Volume 16, Issue 3, Pages 261-273, https://doi.org/10.1016/j.eij.2015.06.005.

[3] Cami BR, Hassanpour H, Mashayekhi H (2018) A content-based movie recommender system based on temporal user preferences. Proc - 3rd Iran Conf Signal Process Intell Syst ICSPIS 2017 2017-Decem:121–125. https://doi.org/10.1109/ICSPIS.2017.8311601.

[4] Chen HW, Wu YL, Hor MK, Tang CY (2017) Fully content-based movie recommender system with feature extraction using neural network. Proc 2017 Int Conf Mach learn Cybern ICMLC 2017 2:504–509. https://doi.org/10.1109/ICMLC.2017.8108968.

[5] Wibowo, Kurnia & Baizal, Z. (2022). Movie Recommendation System Using Knowledge-Based Filtering and K-Means Clustering. Building of Informatics, Technology and Science (BITS), vol.3, Issue 4, PP. 460-465. https://doi.org/10.47065/bits.v3i4.1236.

[6] H. El Bouhissi, M. Adel, A. Ketam, and A. B. M. Salem, “Towards an efficient knowledge-based recommendation system,” CEUR Workshop Proc., vol. 2853, pp. 38–49, 2021.

[7] Jain KN, Kumar V, Kumar P, Choudhury T (2018) Movie recommendation system: hybrid information filtering system. Adv Intell Syst Comput 673:677–686. https://doi.org/10.1007/978-981-10-7245-1_66.

[8] Lekakos G, Caravelas P (2008) A hybrid approach for movie recommendation. Multimed Tools Appl 36:55–70. https://doi.org/10.1007/s11042-006-0082-7.

[9] Thakker, U., Patel, R. & Shah, M. (2021), A comprehensive analysis on movie recommendation system employing collaborative filtering. Multimed Tools Appl 80, 28647–28672. https://doi.org/10.1007/s11042-021-10965-2

[10] Aggarwal CC, Aggarwal CC (2016) Content-based recommender systems. Recomm Syst 139–166. https://doi.org/10.1007/978-3-319-29659-3_4

[11] G. Adomavicius, and A. Tuzhilin (2005),. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), pp. 734–749. https://doi.org/10.1109/TKDE.2005.99

[12] N. Immaneni, I. Padmanaban, B. Ramasubramanian and R. Sridhar. (2017), "A meta-level hybridization approach to personalized movie recommendation," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2193-2200, https://doi.org/10.1109/ICACCI.2017.8126171

[13] R. Lavanya, U. Singh and V. Tyagi. (2021), "A Comprehensive Survey on Movie Recommendation Systems," International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 532-536, https://doi.org/10.1109/ICAIS50930.2021.9395759