International Journal of Informatics and Data Science
Volume 2, No 2, June 2025, Page: 46−54
ISSN 3026-7315 (media online)
https://journals.adaresearch.or.id/ijids/index
DOI: doi.org/10.64366/ijids.v2i2.81

Film Popularity Analysis through Combined K-Means Clustering and Gradient Boosted Trees

Agi Candra Bramantia[1], Desyanti[2], Jeperson Hutahaean[3], Erlin Windia Ambarsari[1,*]

[1] Faculty of Engineering and Computer Science, Informatics Engineering, Universitas Indraprasta PGRI, DKI Jakarta, Indonesia
[2] Informatics Engineering, Sekolah Tinggi Teknologi Dumai, Riau, Indonesia
[3] Faculty of Computer Science, Information System, Universitas Royal, Asahan, Sumatera Utara, Indonesia
Email: [1]agibramantia@yahoo.com, [2]desyanti734@gmail.com, [3]jepersonhutahean@gmail.com, [4,*]erlinunindra@gmail.com
Correspondence Author Email: erlinunindra@gmail.com

Abstract–The dynamic and competitive nature of the global film industry presents complex challenges in predicting film popularity, as success is shaped by the interplay of production investment, casting decisions, and audience preferences. This research addresses the limitations of previous studies that have focused primarily on direct relationships, such as budget versus box office returns, by introducing an integrated analytical framework that combines K-Means clustering and Gradient Boosted Trees (GBT) with explainable AI techniques. Utilizing the TMDB movie dataset and constructing features such as actor influence and studio power, the study segments films and predicts audience ratings while providing interpretable visualizations. The results reveal four distinct film clusters and demonstrate that actor influence and budget allocation are the most significant predictors of popularity. The proposed model achieves an R^2 score of 0.75 and a mean squared error of 0.35 in predicting audience ratings, while cluster analysis shows that Blockbuster films reach the highest average ratings (6.76), and Underperforming films the lowest (2.42). By integrating interpretable predictive modeling and interactive scenario tools, this research offers both theoretical advancement and practical value for industry stakeholders. However, the findings are limited by the available metadata and do not account for factors such as marketing or real-time audience trends, suggesting opportunities for future research to expand the analytical framework.

Keywords: Film Popularity; Machine Learning; Clustering; Explainable AI; Audience Ratings

1. INTRODUCTION

The global film industry is among the most dynamic and influential sectors in the entertainment domain, generating thousands of new films each year across a wide range of cultural and economic contexts. These productions differ in terms of scale, budget allocation, cast composition, and intended audience. In today's competitive market, film success is measured not only by box office revenue, but also by popularity indicators such as audience ratings, online reviews, and digital engagement.

Contrary to common assumptions, large budgets or the involvement of high-profile actors do not guarantee a film's commercial or critical success. Many high-budget films have failed to attract significant attention, while some low-budget productions featuring lesser-known actors have achieved remarkable popularity. These patterns reveal the complexity of factors influencing film success and highlight the importance of systematic, data-driven investigation.

The advent of large-scale online film databases, such as The Movie Database (TMDB), has enabled researchers to explore comprehensive metadata, including production budgets, cast details, studio affiliations, and audience feedback, for films produced worldwide. While several studies have leveraged such data, most have been limited to examining direct relationships—most notably between budget and box office returns—without delving into more complex, multidimensional interactions among production and cast variables.

Recent research has contributed valuable approaches in this domain. Sarker et al.[1] utilized K-Means clustering to segment films according to features such as budget, popularity, and duration, yet did not address predictive modeling. Alzakan et al.[2] enhanced clustering outcomes by applying Gradient Boosting as a post-processing step, improving clustering accuracy and robustness. Leem et al.[3] introduced DRECE, a framework that combines dimensionality reduction, clustering, and explainable classification for box office analysis, with cluster labels included as classification features. Lu et al. [4] developed a movie box office prediction model using a Generalized Regression Neural Network (GRNN) optimized by an Improved Fruit Fly Optimization Algorithm, demonstrating superior prediction performance over conventional methods. Tang[5] validated the use of optimized XGBoost algorithms for box office forecasting, with a focus on marketing and distribution data.

Despite these advances, previous studies typically treat clustering and predictive modeling as separate processes. The explicit integration of cluster-based segmentation as an input feature within a supervised learning framework remains rare. Additionally, there is a need for research that visualizes cluster structures to interpret the joint influence of production and casting attributes on film popularity.

To address these gaps, this study introduces an integrated analytical framework that combines K-Means clustering and Gradient Boosted Trees (GBT) to analyze TMDB film metadata. In the proposed approach, films are initially grouped based on key production characteristics. The resulting cluster labels are then incorporated into a GBT model to predict film popularity, with visualization techniques applied to interpret further the cluster structures concerning individual film features.

International Journal of Informatics and Data Science
Volume 2, No 2, June 2025, Page: 46−54
ISSN 3026-7315 (media online)
https://journals.adaresearch.or.id/ijids/index
DOI: doi.org/10.64366/ijids.v2i2.81

By synthesizing production factors such as budget, studio reputation, and cast experience within a unified analysis pipeline, this research aims to offer practical insights for industry stakeholders, including producers and marketing professionals, while contributing to the ongoing development of data-driven strategies in the creative industries. This research illustrates how visual and explainable machine learning can be effectively used to analyze the combined impact of production and casting attributes on film popularity.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This research was conducted in a structured sequence to ensure methodological rigor and reproducibility. The workflow is illustrated in Figure 1:

Figure 1. Research workflow for cluster-based film popularity prediction.

As illustrated in Figure 1, the research workflow unfolds through a series of stages, each of which is explored in detail below.

a. Data Collection and Preparation: The TMDB dataset [6] was obtained, containing film-level metadata such as production budget, cast, studio, and audience ratings. Data cleaning was conducted, including merging datasets, handling missing values, parsing nested columns, standardizing features, and addressing outliers, in line with recent best practices [7], [8], [9].
b. Feature Engineering: New explanatory variables such as actor_influence (the average historical rating of top actors) and studio_power (the average rating for a main production studio) were constructed based on established methods for feature construction and machine learning in movie prediction tasks [10], [11]. The actor_budget_synergy feature was introduced to model the interaction between cast performance and production scale, drawing inspiration from multi-view user modeling approaches [12].
c. Data Normalization: Standardization using StandardScaler [13] was performed after dataset merging and feature engineering, as recommended in recent movie analytics literature [14], to ensure all input features were on a comparable scale for clustering and predictive modeling.
d. Clustering (K-Means): K-Means clustering was performed on standardized features, and the optimal number of clusters was determined using both the elbow method and the silhouette method. The effectiveness of K-Means for segmenting movie datasets, as well as the use of silhouette analysis for cluster validation, has been demonstrated in several recent works [6], [15], [16].
e. Visualization and Dimensionality Reduction: In this study, cluster assignments from K-Means were visualized using the t-distributed stochastic neighbor embedding (t-SNE) algorithm [17], [18], which reduces the high-dimensional feature space into two dimensions to facilitate interpretation of cluster structures. The adoption of explainable AI techniques—such as SHAP (SHapley Additive exPlanations) for feature importance and t-SNE for visualizing latent patterns—has become increasingly important in the field of media analytics, as these methods provide greater transparency and insight into complex models and data relationships [19]. The impact of t-SNE visualizations will be demonstrated and interpreted in the results section
f. Supervised Learning (GBT/XGBoost): XGBoost-based Gradient Boosted Trees [20], [21] were trained to predict film popularity, using both original and cluster-based features. The choice of this algorithm was motivated by its superior performance and scalability for structured, high-dimensional datasets common in movie analytics.
g. Evaluation: Model performance was measured using Mean Squared Error (MSE), which penalizes large prediction errors, and R^2, which reflects the model's ability to explain variance in the target variable. The use of both metrics offers a balanced view of absolute prediction accuracy and overall model fit [22], [23].
h. Interpretation: To provide transparent explanations of model predictions, SHAP values were employed, offering consistent and intuitive measures of feature contributions at both the global and individual levels [24]. This interpretability aligns with best practices in modern machine learning for media analytics.

2.2 Methods and Algorithmic Foundation

This study used a hybrid of unsupervised and supervised machine learning for movie analytics:

a. K-Means Clustering: A partition-based clustering algorithm, K-Means assigns data points to clusters by minimizing within-cluster variance [25]:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - u_i\|^2 \quad (1)$$

With J as total within-cluster sum of squares and μ_i as cluster centroids. The silhouette coefficient used to assess clustering validity is defined as:

Volume 2, No 2, June 2025, Page: 46−54
ISSN 3026-7315 (media online)
https://journals.adaresearch.or.id/ijids/index
DOI: doi.org/10.64366/ijids.v2i2.81

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (2)$$

The silhouette score uses $a(i)$ as the average distance within a cluster and $b(i)$ as the distance to the nearest cluster, offering a balance of cohesion and separation. This metric is widely used to assess clustering quality and has shown practical value in film data segmentation tasks [26].

b. Gradient Boosted Trees (GBT/XGBoost): An ensemble regression method that sequentially fits multiple decision trees, progressively minimizing prediction errors from previous stages, which significantly improves predictive accuracy and robustness [27], [28]:

$$y_i = \sum_{m=1}^{M} f_m(x_i) \quad (3)$$

Where f_m denotes the m-th tree and M is the total number of boosting rounds.

c. SHAP (SHapley Additive exPlanations): SHAP provides a game-theoretic approach to interpret model predictions by attributing contributions of each feature. The SHAP value for feature j is defined as [29], [30]:

$$\phi_j = \sum_{s \subseteq F \setminus \{j\}} \frac{|s|!(|F|-|s|-1)!}{|F|!} [fs \cup \{j\}(xs \cup \{j\}) - fs(xs)] \quad (4)$$

Where F is the set of all features, S is a subset of features not containing j, and f_S is the model trained with features in S.

d. t-SNE Visualization: t-distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction technique that visualizes high-dimensional data in two or three dimensions. The use of t-SNE for visualization is increasingly recommended for high-dimensional datasets in media and film. Table 1 lists the primary features incorporated into the clustering and prediction processes. All stages and algorithms were implemented in Python using the Scikit-learn and XGBoost libraries

Table 1. Main features derived from TMDB data.

Feature	Description
budget	Total production budget (USD)
runtime	Duration of the film (minutes)
actor_influence	Mean rating of top actors' previous films
studio_power	Average rating of films by the main production studio
actor_budget_synergy	Product of actor_influence and budget
cluster_label	Cluster assignment (K-Means)

3. RESULT AND DISCUSSION

The present analysis employed a combined approach of clustering, supervised learning, and model interpretability to elucidate the factors contributing to film popularity. Each step of the process is reflected in direct outputs and visualizations, allowing for a comprehensive understanding of how production choices and casting interact in shaping audience reception.

3.1 Film Segmentation through Clustering

The application of K-Means clustering, followed by t-SNE visualization, revealed clear segmentation within the film dataset. As illustrated in Figure 2, films naturally grouped into four distinct clusters when considering attributes such as budget, runtime, actor influence, and studio power.

International Journal of Informatics and Data Science
Volume 2, No 2, June 2025, Page: 46−54
ISSN 3026-7315 (media online)
https://journals.adaresearch.or.id/ijids/index
DOI: doi.org/10.64366/ijids.v2i2.81

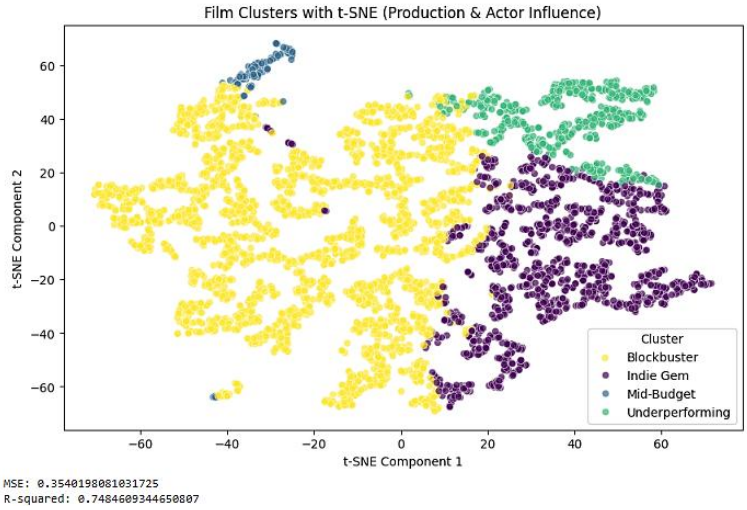

Figure 2. t-SNE visualization of K-Means clusters, representing film segmentation based on key production and cast features.

The t-SNE plot in Figure 2 clearly distinguishes the four main clusters: Blockbuster, Indie Gem, Mid-Budget, and Underperforming. In this dataset, Indie Gems builds the largest group with 2,854 films, followed by Blockbusters with 1,328 films, Mid-Budget with 480 films, and Underperforming with 139 films. This distribution highlights not only the dominance of independent productions but also the diversity of resource allocation strategies present in the film industry.

The positioning of clusters on the t-SNE map reflects genuine differentiation in creative and resource profiles. The Blockbuster, Indie Gem, Mid-Budget, and Underperforming segments each display unique patterns, indicating that industry structure extends beyond a simple high-budget versus low-budget dichotomy. The emergence of these clusters affirms that both production scale and talent configuration play pivotal roles in determining the character and market trajectory of a film. This visualization directly addresses the research objective by demonstrating that film segmentation based on both production and casting attributes uncovers meaningful patterns, which were not captured by prior studies focusing only on single factors.

This segmentation not only demonstrates the presence of distinct film groups, but also fulfills the main research objective to uncover multidimensional patterns in film data that are often overlooked in traditional one-dimensional analyses.

3.2 Drivers of Film Ratings: Model-Based Feature Importance

Subsequent to clustering, the study utilized Gradient Boosted Trees (XGBoost) to predict film ratings based on the previously identified features. SHAP analysis, as visualized in Figure 3, provides an interpretable summary of how each variable influences the model's predictions.

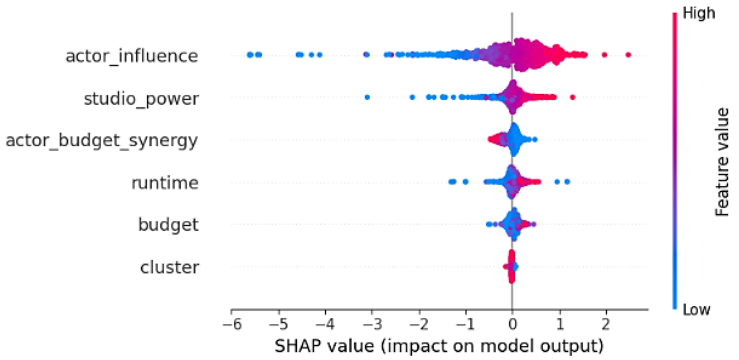

Figure 3. SHAP summary plot depicting the relative influence of each input feature on predicted film ratings.

The SHAP analysis in Figure 3 ranks actor influence as the most impactful feature for predicting film ratings, followed by studio power and actor-budget synergy. This ranking quantitatively confirms that both talent reputation and production investment—especially when combined—significantly shape audience response.

Actor influence emerged as the most significant contributor, reinforcing the established understanding that casting decisions—particularly involving high-profile or proven talent—tend to sway audience reception. The importance of studio power and the interaction between actor influence and budget (actor-budget synergy) further demonstrate that successful outcomes are rarely attributable to a single dimension. These results extend the findings

International Journal of Informatics and Data Science
Volume 2, No 2, June 2025, Page: 46−54
ISSN 3026-7315 (media online)
https://journals.adaresearch.or.id/ijids/index
DOI: doi.org/10.64366/ijids.v2i2.81

of Sarker et al. [1] and Leem et al. [3], suggesting that the intersection of creative and financial resources constitutes a major determinant of popularity.

3.3 Explaining Individual Model Predictions

To enhance interpretability at the level of individual films, the SHAP waterfall plot in Figure 4 demonstrates how specific feature values contribute to the final predicted rating.

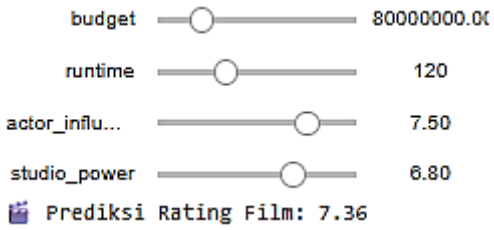

Figure 4. SHAP waterfall plot illustrating the stepwise contribution of each feature to a single film's predicted rating.

The waterfall plot displays the baseline average rating and the incremental effect of each variable, enabling a transparent audit of the model's logic. Positive contributions from actor influence and studio power are visible in this case, while other features may offset or reinforce the prediction, depending on their values. Providing this level of detailed, instance-based explanation ensures that the modeling approach is not only predictive but also actionable for decision-makers, as targeted in the research objectives.

3.4 Practical Model Application: Interactive Scenario Analysis

The deployment of an interactive prediction tool (Figure 5) allows practitioners to simulate outcomes for different combinations of film attributes. Users can adjust budget, runtime, actor influence, and studio power, observing both the predicted rating and an explanation of feature contributions in real-time.

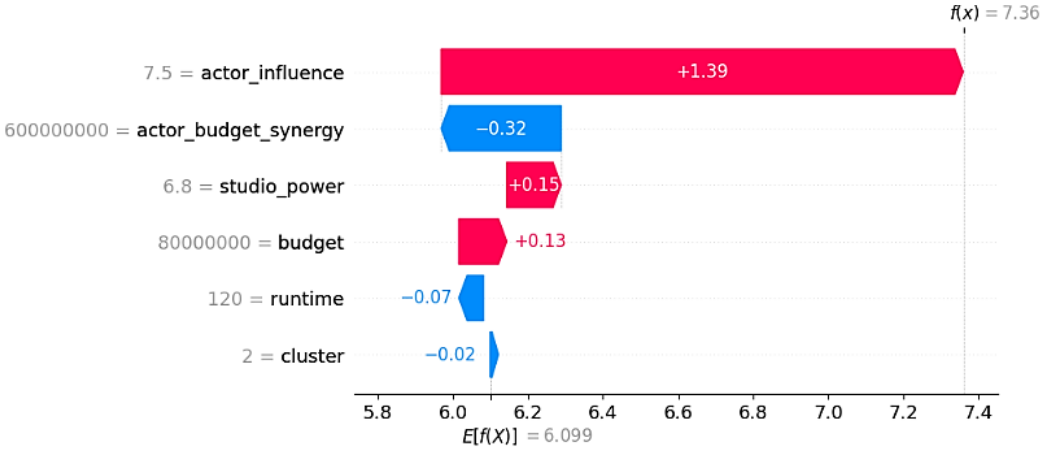

Figure 5. Interactive prediction tool for film ratings, integrating scenario-based analysis and SHAP explanations.

Figure 5 displays a SHAP waterfall plot for a single film prediction. The baseline average rating of 6.10 is incrementally adjusted by each feature: actor influence (7.5) increases the predicted rating by +1.39, actor-budget synergy decreases it by -0.32, while studio power (6.8) and budget ($80 million) add smaller positive effects. Runtime and cluster assignment provide minor negative adjustments. The model's final prediction for this scenario is 7.36.

This breakdown makes it clear how individual casting and production decisions—especially actor influence—impact audience ratings. For example, an increase in actor influence translates directly to a higher predicted rating, indicating its strategic importance for producers. By providing such granular explanations, the model moves beyond black-box predictions and becomes a useful tool for real-world scenario analysis and resource planning.

This practical interpretability is central to the study's aim of equipping industry stakeholders with actionable, data-driven support for their decision-making processes, as emphasized in the research objectives.

3.5 Statistical Profiles of Clusters

Direct statistical profiling of each cluster is presented in Figure 6, detailing the number of films, average audience rating, production budget, and runtime.

International Journal of Informatics and Data Science
Volume 2, No 2, June 2025, Page: 46−54
ISSN 3026-7315 (media online)
https://journals.adaresearch.or.id/ijids/index
DOI: doi.org/10.64366/ijids.v2i2.81

```
--- Statistik per Cluster ---
     cluster_label     N  Avg_Rating  Std_Dev_Rating  Avg_Budget_M  \
0      Blockbuster  1328    6.758735        0.734795     25.485775
1        Indie Gem  2854    5.928241        0.995483     15.549159
2       Mid-Budget   480    6.295417        0.776091    126.849786
3  Underperforming   139    2.423022        1.753271      2.719579

   Std_Dev_Budget_M  Avg_Runtime
0         22.323899   128.500000
1         19.098171    95.871409
2         46.244982   118.714583
3          9.056971    85.345324
```

Figure 6. Summary statistics for each cluster, capturing volume, ratings, budget, and runtime.

Figure 6 summarizes the main characteristics of each cluster. Blockbusters record the highest average audience rating at 6.76 with relatively moderate budgets averaging $25.5 million and longer runtimes of 128.5 minutes. In contrast, Mid-Budget films invest significantly more, averaging $126.8 million per film, but achieve a slightly lower mean rating of 6.30. Indie Gems, while the most numerous, maintain a moderate average rating of 5.93 and operate with budgets around $15.5 million. Underperforming films stand out with the lowest ratings (2.42) and minimal budgets ($2.7 million), underscoring the challenges of films produced with limited resources.

Blockbusters display the highest average ratings and extended runtimes, though Mid-Budget films allocate greater resources on average. Indie Gems are notable for their volume and diversity, while Underperforming films, as expected, report both the lowest ratings and most limited budgets. The standard deviations indicate considerable heterogeneity within clusters, particularly among Indie Gems, suggesting the presence of both niche successes and outliers.

3.6 Comparative Visualization of Cluster Characteristics

To facilitate comparison across clusters, Figures 7 to 9 present bar charts summarizing average audience ratings, production budgets, and runtimes.

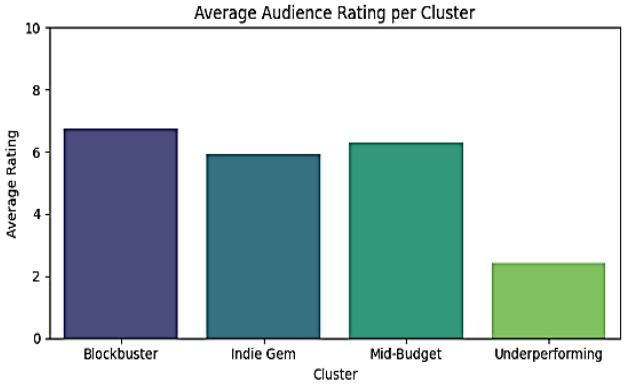

Figure 7. Blockbusters are associated with the highest mean audience ratings, followed by Mid-Budget productions. Indie Gems and Underperforming films occupy lower tiers.

Figure 7 illustrates the comparative average audience ratings across clusters. Blockbusters lead with an average of 6.76, followed by Mid-Budget films at 6.30, Indie Gems at 5.93, and Underperforming films at just 2.42. This clear gap between segments reinforces the notion that higher ratings are not always tied to the largest budgets, and that production and casting strategies play a pivotal role in audience reception.

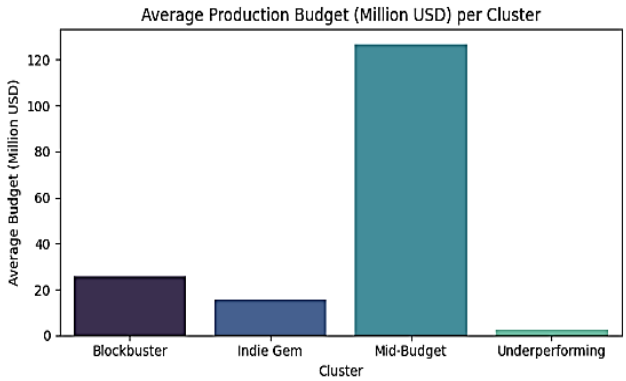

Figure 8. Mid-Budget films demonstrate the highest average production budgets, suggesting a strategic focus on investment, while Indie Gems and Underperforming clusters are more constrained financially.

International Journal of Informatics and Data Science
Volume 2, No 2, June 2025, Page: 46−54
ISSN 3026-7315 (media online)
https://journals.adaresearch.or.id/ijids/index
DOI: doi.org/10.64366/ijids.v2i2.81

As shown in Figure 8, Mid-Budget films allocate the largest budgets, averaging $126.8 million, which is five times greater than the average Blockbuster budget. Nevertheless, this investment does not guarantee the highest ratings, as seen in the earlier comparison, indicating the importance of effective resource utilization over absolute spending.

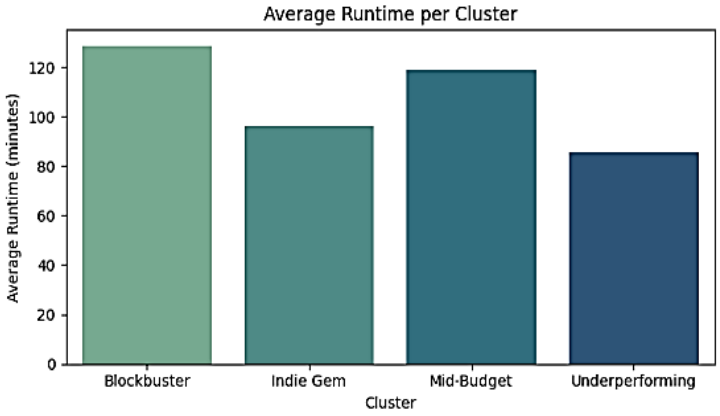

Figure 9. Blockbusters consistently have the longest runtimes, supporting their characterization as immersive, large-scale projects.

Figure 9 presents the average runtimes across clusters, with Blockbusters averaging 128.5 minutes, Mid-Budget films at 118.7 minutes, Indie Gems at 95.9 minutes, and Underperforming films at 85.3 minutes. The longer runtimes of Blockbusters reflect their positioning as feature-length, large-scale productions intended for wide release. For instance, even though Mid-Budget films invest five times more than Blockbusters, their mean rating remains lower. This suggests that efficient resource allocation and casting decisions may be more valuable than simply increasing budgets.

These visualizations confirm that production investment, creative ambition, and popularity are interrelated but not strictly dependent on one another. In particular, high expenditure does not always result in high ratings, and the largest cluster by count (Indie Gems) demonstrates that diversity and experimentation remain central to the industry.

3.7 Integrating Findings with Industry Context

The synthesis of clustering, predictive modeling, and explainable AI yields a nuanced understanding of film success. Blockbusters stand out for consistent, high ratings and substantial runtimes, often benefiting from recognizable talent and strong production houses. Mid-Budget films represent a significant investment tier but may not always match Blockbusters in audience acclaim. Indie Gems reflect the dynamic, risk-taking segment of the industry, offering both innovation and unpredictability. Underperforming films, while smaller in both scale and impact, contribute to the broader diversity of cinematic output. Such segmentation and explainable modeling provide clear, data-driven guidance for producers in making investment and casting decisions that align with market realities.

These patterns underscore that there are multiple, viable strategies for film production. The findings also support the introduction's premise that popularity is not guaranteed by budget alone; rather, it is shaped by the interplay of creative and production choices.

3.8 Implications and Future Research

By providing a transparent, reproducible analytical process and direct model explanations, this research supports both strategic planning and knowledge transfer within the film industry. The integration of scenario analysis and feature interpretation enables decision-makers to evaluate options more effectively and respond to market complexity with greater confidence.

Further studies may expand the feature set, incorporate additional sources of feedback (such as critical reviews or audience demographics), or explore alternative clustering and prediction methods to refine segmentation and improve forecasting.

4. CONCLUSION

This research demonstrates that integrating K-Means clustering and Gradient Boosted Trees, supported by explainable AI visualizations, enables a more comprehensive understanding of how production factors and casting attributes jointly shape film popularity. By visualizing cluster-based patterns and identifying the dominant influence of actor reputation and resource allocation, the study provides evidence that success in the film industry is determined by a combination of creative and financial strategies rather than by singular factors such as budget or star power alone. The use of interactive predictive tools further extends the practical value of these findings for

Volume 2, No 2, June 2025, Page: 46−54
ISSN 3026-7315 (media online)
https://journals.adaresearch.or.id/ijids/index
DOI: doi.org/10.64366/ijids.v2i2.81

industry stakeholders, allowing scenario-based planning and more transparent decision support. However, the study is subject to several limitations. The analysis relies exclusively on the TMDB dataset, which, while extensive, may not fully capture external influences such as marketing campaigns, distribution networks, or evolving audience behavior. Additionally, the dataset represents a snapshot in time, which restricts the ability to assess long-term shifts or emerging trends in the industry. The chosen methods, although effective in uncovering correlations, cannot fully disentangle causal relationships between creative and financial variables. Future studies are encouraged to incorporate additional contextual factors, longitudinal data, and complementary modeling approaches to refine predictive accuracy and provide a more holistic understanding of film performance.

REFERENCES

[1] K. U. Sarker *et al.*, "A Ranking Learning Model by K-Means Clustering Technique for Web Scraped Movie Data," *Computers*, vol. 11, no. 11, Nov. 2022, doi: 10.3390/computers11110158.

[2] M. Alzakan, H. Almousa, A. Almarzoqi, M. Alghasham, M. Aldawsari, and M. Al-Hagery, "Enhancing K-means Clustering Results with Gradient Boosting: A Post-Processing Approach," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, 2024, doi: 10.14569/IJACSA.2024.0150292.

[3] S. Leem, J. Oh, D. So, and J. Moon, "Towards Data-Driven Decision-Making in the Korean Film Industry: An XAI Model for Box Office Analysis Using Dimension Reduction, Clustering, and Classification," *Entropy*, vol. 25, no. 4, p. 571, Mar. 2023, doi: 10.3390/e25040571.

[4] W. Lu, X. Zhang, and X. Zhan, "Movie Box Office Prediction Based on IFOA-GRNN," *Discrete Dyn Nat Soc*, vol. 2022, no. 1, p. 3690077, Jan. 2022, doi: https://doi.org/10.1155/2022/3690077.

[5] S. Tang, "The box office prediction model based on the optimized XGBoost algorithm in the context of film marketing and distribution," *PLoS One*, vol. 19, no. 10, p. e0309227, Oct. 2024, doi: 10.1371/journal.pone.0309227.

[6] I. F. Ashari, R. Banjarnahor, D. R. Farida, S. P. Aisyah, A. P. Dewi, and N. Humaya, "Application of Data Mining with the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies," *Journal of Applied Informatics and Computing*, vol. 6, no. 1, pp. 07–15, Jul. 2022, doi: 10.30871/jaic.v6i1.3485.

[7] C. Xie, "A refined approach to early movie box office prediction leveraging ensemble learning and feature encoding," *Applied and Computational Engineering*, vol. 75, no. 1, pp. 273–284, Jul. 2024, doi: 10.54254/2755-2721/75/20240555.

[8] Y. Zheng, "Predicting Movie Box Office Based on Machine Learning, Deep Learning, and Statistical Methods," *Applied and Computational Engineering*, vol. 94, no. 1, pp. 20–32, Oct. 2024, doi: 10.54254/2755-2721/94/2024MELB0069.

[9] A. Singh, P. Singh, and A. K. Tiwari, "A Comprehensive Survey on Machine Learning," *Journal of Management and Service Science (JMSS)*, vol. 1, no. 1, pp. 1–17, Mar. 2021, doi: 10.54060/JMSS/001.01.003.

[10] Y. Zheng, "Predicting Movie Box Office Based on Machine Learning, Deep Learning, and Statistical Methods," *Applied and Computational Engineering*, vol. 94, no. 1, pp. 20–32, Oct. 2024, doi: 10.54254/2755-2721/94/2024MELB0069.

[11] S. Çağlıyor, B. Öztayşi, and S. Sezgin, "Forecasting Box Office Performances Using Machine Learning Algorithms," in *Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making*, C. Kahraman, S. Cebi, S. Cevik Onar, B. Oztaysi, A. C. Tolga, and I. U. Sari, Eds., Cham: Springer International Publishing, 2020, pp. 257–264.

[12] S. Li, R. Xie, Y. Zhu, X. Ao, F. Zhuang, and Q. He, "User-Centric Conversational Recommendation with Multi-Aspect User Modeling," in *SIGIR 2022 - Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, Jul. 2022, pp. 223–233. doi: 10.1145/3477495.3532074.

[13] K. M. Sujon, R. B. Hassan, Z. T. Towshi, M. A. Othman, M. A. Samad, and K. Choi, "When to Use Standardization and Normalization: Empirical Evidence From Machine Learning Models and XAI," *IEEE Access*, vol. 12, pp. 135300–135314, 2024, doi: 10.1109/ACCESS.2024.3462434.

[14] N. Pavitha *et al.*, "Movie recommendation and sentiment analysis using machine learning," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 279–284, 2022, doi: https://doi.org/10.1016/j.gltp.2022.03.012.

[15] A. G., S. S. Rao, and K. Chandrasekaran, "Application of Machine Learning in Movie Recommendation using Harris Hawks Optimization and K-means (HHO-k-means) Clustering," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 7s, pp. 515–525, Jul. 2023, [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/2990

[16] T. Widiyaningtyas, I. Hidayah, and T. B. Adji, "Recommendation algorithm using clustering-based upcsim (Cb-upcsim)," *Computers*, vol. 10, no. 10, Oct. 2021, doi: 10.3390/computers10100123.

[17] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu, "Revisiting Dimensionality Reduction Techniques for Visual Cluster Analysis: An Empirical Study," *IEEE Trans Vis Comput Graph*, vol. 28, no. 01, pp. 529–539, 2022, doi: 10.1109/TVCG.2021.3114694.

Volume 2, No 2, June 2025, Page: 46−54
ISSN 3026-7315 (media online)
https://journals.adaresearch.or.id/ijids/index
DOI: doi.org/10.64366/ijids.v2i2.81

[18] T. T. Cai and R. Ma, "Theoretical foundations of t-SNE for visualizing high-dimensional clustered data," *J. Mach. Learn. Res.*, vol. 23, no. 1, Jan. 2022.
[19] S. A. and S. R., "A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends," *Decision Analytics Journal*, vol. 7, p. 100230, 2023, doi: https://doi.org/10.1016/j.dajour.2023.100230.
[20] S. Tang, "The box office prediction model based on the optimized XGBoost algorithm in the context of film marketing and distribution," *PLoS One*, vol. 19, no. 10, pp. e0309227-, Oct. 2024, [Online]. Available: https://doi.org/10.1371/journal.pone.0309227
[21] Y. Filmus, I. Mehalel, and S. Moran, "A Resilient Distributed Boosting Algorithm," in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., in Proceedings of Machine Learning Research, vol. 162. PMLR, Sep. 2022, pp. 6465–6473. [Online]. Available: https://proceedings.mlr.press/v162/filmus22a.html
[22] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, p. e623, 2021, doi: 10.7717/peerj-cs.623.
[23] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," *Frontiers in Bioinformatics*, vol. Volume 4-2024, 2024, [Online]. Available: https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2024.1457619
[24] S. J. Silva, C. A. Keller, and J. Hardin, "Using an Explainable Machine Learning Approach to Characterize Earth System Model Errors: Application of SHAP Analysis to Modeling Lightning Flash Occurrence," *J Adv Model Earth Syst*, vol. 14, no. 4, p. e2021MS002881, Apr. 2022, doi: https://doi.org/10.1029/2021MS002881.
[25] Z. Ning, J. Chen, J. Huang, U. J. Sabo, Z. Yuan, and Z. Dai, "WeDIV – An improved k-means clustering algorithm with a weighted distance and a novel internal validation index," *Egyptian Informatics Journal*, vol. 23, no. 4, pp. 133–144, 2022, doi: https://doi.org/10.1016/j.eij.2022.09.002.
[26] P. Bombina, D. Tally, Z. B. Abrams, and K. R. Coombes, "SillyPutty: Improved clustering by optimizing the silhouette width," *PLoS One*, vol. 19, no. 6, p. e0300358, Jun. 2024, doi: 10.1371/journal.pone.0300358.
[27] E. D. Omar *et al.*, "Comparative Analysis of Logistic Regression, Gradient Boosted Trees, SVM, and Random Forest Algorithms for Prediction of Acute Kidney Injury Requiring Dialysis After Cardiac Surgery," *Int J Nephrol Renovasc Dis*, vol. 17, pp. 197–204, Jul. 2024, doi: 10.2147/IJNRD.S461028.
[28] L. W. Rizkallah, "Enhancing the performance of gradient boosting trees on regression problems," *J Big Data*, vol. 12, no. 1, p. 35, 2025, doi: 10.1186/s40537-025-01071-3.
[29] X. Huang *et al.*, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Comput Sci Rev*, vol. 37, p. 100270, 2020, doi: https://doi.org/10.1016/j.cosrev.2020.100270.
[30] A. T. Keleko, B. Kamsu-Foguem, R. H. Ngouna, and A. Tongne, "Health condition monitoring of a complex hydraulic system using Deep Neural Network and DeepSHAP explainable XAI," *Advances in Engineering Software*, vol. 175, p. 103339, 2023, doi: https://doi.org/10.1016/j.advengsoft.2022.103339.