ISSN 2089-385X (Print) | 2829-6761 (Online), Volume 15, No. 2 (2026), pp. 18-24

Vertex

Published by: Institute of Computer Science (IOCS)

IMDb Movie Rating Prediction Using a Random Forest Classification Approach

Rifqy Rosyidah Ilmi
Informatika Kesehatan, Universitas Sunan Gresik, Indonesia

Abstract

Accurate movie rating prediction is essential for supporting audience preferences and analytical decision-making in the digital film industry. The availability of large-scale metadata from IMDb provides valuable opportunities for applying machine learning techniques to analyze rating patterns. This study investigates the effectiveness of a Random Forest classification model for predicting IMDb movie rating categories based on structured attributes, including genre, movie duration, content rating, actor popularity, and user review statistics. Data preprocessing involved handling missing values, removing duplicates, encoding categorical variables, normalizing numerical features, and partitioning the dataset into training and testing subsets. To mitigate class imbalance among rating categories, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. Experimental evaluation demonstrates that the proposed model achieves an overall accuracy of 0.78, accompanied by balanced precision, recall, and F1-score values across all classes. Confusion matrix analysis shows that classification errors predominantly occur between neighboring rating categories, reflecting the inherent subjectivity of movie ratings. Furthermore, feature importance analysis highlights genre, duration, content rating, and user engagement indicators as the most influential predictors. These results indicate that Random Forest offers a robust and interpretable baseline model for IMDb rating prediction and provides meaningful insights for future movie analytics and recommendation research.

Article Info

Article history:

Received : Jan 06,2026
Revised : Jan 23, 2026
Accepted : Feb 02, 2026

Keywords:

IMDb;
Machine Learning;
Random Forest;
SMOTE

Corresponding Author:

Rifqy Rosyidah Ilmi,
Informatika Kesehatan,
Universitas Sunan Gresik,
Jl. KH. Syafi'i No. 15 Dahanrejo Kebomas Gresik Jawa Timur 61124.
rifqy.ri@lecturer.usg.ac.id.

Introduction

Online movie rating services have become an important digital reference for audiences when selecting films and for industry stakeholders when evaluating market performance. Among these services, the Internet Movie Database (IMDb) provides extensive information, including ratings, reviews, and detailed movie attributes, which enables large-scale analytical studies. Movie ratings influence not only viewer decisions but also promotional strategies, revenue forecasting, and the development of recommendation systems (Ahmad & Khalid, 2021; Liu et al., 2022).

However, IMDb ratings inherently reflect subjective human perceptions and tend to form uneven distributions, where moderate rating categories dominate while extreme ratings appear less frequently. This imbalance poses a significant challenge for supervised learning models, as classifiers trained on skewed data tend to be biased toward majority classes, leading to reduced predictive performance and unreliable decision boundaries, particularly for minority rating categories. Therefore, addressing class imbalance is crucial to ensure fair, stable, and generalizable classification outcomes, especially in the context of subjective movie rating data.

Recent studies have demonstrated that machine learning techniques can effectively model complex relationships between movie attributes and audience ratings. Various approaches, including regression models, decision trees, neural networks, and ensemble learning methods, have been applied to movie rating prediction tasks. However, single-model approaches often suffer from limited generalization performance, sensitivity to noise, or low interpretability when applied to heterogeneous movie datasets (Sarker, 2021; Song & Kim, 2020).

Ensemble learning methods, particularly Random Forest, have shown strong performance in classification problems involving structured and high-dimensional data. Random Forest combines multiple decision trees to improve predictive accuracy and reduce overfitting while providing feature importance measures that enhance model interpretability (Breiman, 2001; Wu et al., 2020). Several recent studies have reported the effectiveness of Random Forest in recommendation systems and online platform analytics, especially when dealing with imbalanced class distributions (Wang et al., 2023; Yang et al., 2020).

Despite these methodological advances, existing research predominantly emphasizes predictive performance, while relatively limited attention is devoted to systematically investigating the combined impact of class imbalance handling, model interpretability, and subjective rating characteristics within the IMDb context. Most prior studies apply Random Forest or other classifiers without explicitly examining how oversampling strategies influence model stability, decision boundaries, and feature contribution patterns across different rating categories. From a theoretical perspective, this study contributes by integrating ensemble learning theory, imbalanced learning mechanisms, and explainable machine learning to establish a unified analytical framework for subjective rating prediction. This framework not only enhances classification accuracy under skewed distributions but also provides deeper insights into how intrinsic movie attributes and audience engagement factors jointly shape rating outcomes.

Despite these advances, research focusing specifically on IMDb rating category prediction using Random Forest with comprehensive preprocessing and explicit class imbalance handling remains limited. Many existing studies do not adequately address rating imbalance or lack detailed analysis of model behaviour using confusion matrices and feature importance measures. Therefore, this study investigates the effectiveness of a Random Forest-based classification model for predicting IMDb movie rating categories by incorporating systematic data preprocessing, Synthetic Minority Over-sampling Technique (SMOTE), and feature importance analysis. Specifically, SMOTE is employed to mitigate rating imbalance, stabilize model learning, and ensure equitable representation of minority rating classes, thereby improving robustness and generalization performance.

The key contributions of this study include: (1) establishing a theoretically grounded framework that integrates Random Forest, SMOTE, and explainable learning for subjective rating prediction; (2) providing a comprehensive empirical evaluation of imbalance-aware classification strategies; and (3) delivering interpretable insights into the dominant factors governing IMDb movie ratings, which extend beyond purely performance-driven modelling approaches.

Method

Research Design

This research employs a quantitative approach using supervised machine learning techniques to classify IMDb movie rating categories. The analytical process includes data collection, preprocessing, handling class imbalance, model development using a Random Forest classifier, and performance evaluation. This workflow is consistent with established practices in predictive analytics and recommendation system studies (Liu et al., 2022; Sarker, 2021). The primary objective is to assess the predictive capability and interpretability of the Random Forest model in capturing relationships between movie characteristics and rating outcomes.

Dataset Description

The dataset used in this study was obtained from publicly available IMDb movie metadata repositories, which are widely utilized in movie analytics research due to their richness and reliability(Ahmad & Khalid, 2021; Martínez & Martínez, 2021). The dataset contains structured information on movies, including categorical and numerical attributes such as genre, movie duration, content rating, release year, actor popularity, director popularity, and user engagement indicators (e.g., number of user reviews).

The target variable represents IMDb movie ratings, which are originally continuous values. To formulate a multi-class classification problem, the ratings were discretized into predefined rating categories based on commonly used thresholding strategies in previous studies (Song & Kim, 2020). This transformation allows the model to focus on categorical rating prediction rather than continuous regression.

Data Preprocessing

Data preprocessing was applied to enhance dataset reliability prior to model training. Numerical attributes with missing values were treated using statistical imputation, while missing categorical values were replaced with the most frequent categories. Duplicate entries were eliminated to avoid redundancy. Categorical features were transformed into numerical form through label encoding, enabling compatibility with tree-based algorithms. Although Random Forest models are relatively insensitive to feature scaling, Min–Max normalization was applied to numerical variables to maintain uniform value ranges. The dataset was then split into training and testing sets using an 80:20 ratio to ensure unbiased evaluation of model performance (James et al., 2021).

Class Imbalance Handling

Imbalanced class distributions are commonly observed in IMDb rating datasets, where certain rating categories appear more frequently than others. To mitigate this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented on the training data. SMOTE creates synthetic instances of minority classes by interpolating existing samples, thereby improving class representation without replicating original observations (Chawla et al., 2002). Restricting oversampling to the training set helps prevent information leakage and preserves the validity of performance evaluation (He & Ma, 2021).

Random Forest Model

Random Forest is an ensemble-based learning method that builds multiple decision trees from bootstrapped subsets of the training data, with random feature selection applied at each node split. Predictions are generated through majority voting among the trees, resulting in improved generalization and reduced overfitting (Breiman, 2001). In this study, Random Forest was selected

due to its robustness when handling structured data, resistance to noise, and ability to provide feature importance measures that support model interpretability (Zhou, 2021).

To achieve optimal model performance, hyperparameter tuning was conducted using a grid search strategy combined with k-fold cross-validation. The hyperparameters considered in the grid search included the number of trees (n_estimators), maximum tree depth (max_depth), minimum number of samples required to split an internal node (min_samples_split), and minimum number of samples required at a leaf node (min_samples_leaf). The selection criteria for determining the optimal hyperparameter configuration were based on maximizing the macro-averaged F1-score, which provides a balanced evaluation across all rating classes, particularly under imbalanced data conditions. This metric was chosen instead of accuracy alone to avoid biased optimization toward majority classes and to ensure robust classification performance across minority rating categories. A five-fold cross-validation scheme was employed to ensure stable and reliable performance estimation, where the training data were partitioned into five subsets, iteratively using four folds for training and one fold for validation. The hyperparameter combination yielding the highest mean cross-validated macro F1-score was selected as the final model configuration. Additionally, constraints on maximum tree depth and minimum samples per leaf were applied to prevent overfitting and enhance model generalization capability.

Model Evaluation Metrics

Model effectiveness was assessed using several evaluation metrics. Overall classification accuracy was calculated to measure predictive correctness, while precision, recall, and F1-score were employed to evaluate performance across individual classes, particularly under imbalanced conditions (Saito & Rehmsmeier, 2015). Additionally, a confusion matrix was utilized to examine misclassification patterns between rating categories, offering deeper insight into model behavior (Powers, 2020).

Feature Importance Analysis

To enhance model interpretability, feature importance analysis was conducted based on the impurity reduction mechanism inherent in Random Forest. This analysis quantifies the contribution of each feature to the overall prediction process, allowing identification of dominant factors influencing IMDb movie ratings. Feature importance has been widely used in ensemble-based models to support explainable machine learning and decision-support systems (Lundberg & Lee, 2020; Wang et al., 2023).

Results and Discussion

Classification Performance

The performance of the Random Forest model was evaluated using accuracy, precision, recall, and F1-score to assess its effectiveness in predicting IMDb movie rating categories. As presented in Table 1, the model achieved an overall accuracy of **0.78**, indicating that a substantial proportion of rating instances were correctly classified. The precision value of **0.77** suggests that the model produces relatively few false-positive predictions, while the recall score of **0.76** indicates its ability to identify most relevant instances across rating classes. The resulting F1-score of **0.76** reflects a balanced trade-off between precision and recall, confirming the robustness of the model when applied to imbalanced IMDb rating data (Chawla et al., 2002; He & Ma, 2021).

The application of SMOTE significantly reduced performance disparities between majority and minority rating classes, as evidenced by the relatively balanced precision and recall values across all categories. Compared to preliminary experiments conducted without SMOTE, the minority classes exhibited substantial improvements in recall, indicating enhanced sensitivity

toward underrepresented rating categories. This demonstrates that synthetic oversampling effectively improves class-level predictive fairness without disproportionately inflating majority-class performance.

These results are consistent with recent studies that report strong performance of ensemble-based classifiers for movie rating prediction tasks(Liu et al., 2022; Song & Kim, 2020). The balanced evaluation metrics further indicate that the application of SMOTE effectively mitigates class imbalance without introducing significant bias toward majority classes.

Table 1. Classification Performance of the Random Forest Model

Metric	Value
Accuracy	0,78
Precision	0,77
Recall	0,76
F-1 Score	0,76

Moreover, the stability of the model under synthetic data augmentation is reflected in the low variance observed across cross-validation folds, suggesting that the introduction of synthetic samples does not compromise model generalization or induce instability. This finding supports the robustness of the Random Forest–SMOTE integration for subjective rating classification tasks involving highly skewed distributions.

Confusion Matrix Analysis

Further analysis using a confusion matrix reveals that the majority of predictions are concentrated along the diagonal, indicating a high rate of correct classifications. Misclassifications primarily occur between adjacent rating categories, such as medium and high ratings. This pattern reflects the gradual and subjective nature of audience evaluations, where small perceptual differences may result in similar rating outcomes (Martínez & Martínez, 2021; Powers, 2020). Importantly, the Random Forest model rarely confuses extreme rating categories, suggesting that it successfully captures the overall structure of rating distributions.

The improved classification consistency across minority classes after SMOTE application further confirms that synthetic oversampling enhances inter-class separability without distorting the intrinsic data structure.

Overfitting and Underfitting Analysis

To assess the risks of overfitting and underfitting, model performance was compared between training and testing datasets. The absence of a significant performance discrepancy between training and testing results indicates that the model generalizes well to unseen data without exhibiting excessive variance. This pattern suggests that the model does not overfit the training data and maintains stable predictive behaviour across different data subsets.

Furthermore, the consistently balanced precision, recall, and F1-score values across evaluation sets indicate that the model does not suffer from underfitting, as it successfully captures the underlying relationships between movie attributes and rating outcomes. These results demonstrate that the optimized Random Forest model achieves a stable bias–variance trade-off, ensuring both learning adequacy and generalization capability.

Feature Importance Analysis

Feature importance analysis was conducted to identify the most influential predictors contributing to the Random Forest model's decisions. As shown in Table 2, genre emerges as the most important feature, followed by movie duration and content rating, highlighting the dominant role of content-related attributes in determining movie ratings. User engagement indicators, including actor popularity and number of user reviews, also exhibit substantial importance, underscoring the influence of audience interaction and public visibility on rating outcomes.

These findings align with previous research emphasizing the combined impact of intrinsic movie characteristics and audience-related factors on rating prediction(Ahmad & Khalid, 2021; Wang et al., 2023). The availability of feature importance measures enhances the interpretability of the Random Forest model, making it suitable for decision-support and recommendation system applications where transparency is required (Lundberg & Lee, 2020).

Table 2. Feature Importance Ranking from the Random Forest Model

Rank	Feature	Importance Score
1	Genre	0.24
2	Movie Duration	0.19
3	Content Rating	0.16
4	Actor Popularity	0.14
5	Number of User Reviews	0.13
6	Director Popularity	0.08
7	Release Year	0.06

Notably, the stability of feature importance rankings across cross-validation folds further indicates that the introduction of synthetic samples through SMOTE does not distort the underlying feature–outcome relationships, thereby preserving model interpretability and reliability.

Discussion Summary

The experimental results indicate that the Random Forest classifier performs effectively in predicting IMDb movie rating categories while maintaining interpretability through feature importance analysis. The combination of strong classification metrics and meaningful explanatory insights demonstrates the model's ability to capture complex relationships between movie attributes and audience evaluations. Future studies may extend this approach by incorporating unstructured data sources, such as textual reviews or sentiment-based features, to further improve prediction accuracy.

Conclusion

This study demonstrates that the integration of Random Forest classification, Synthetic Minority Over-sampling Technique (SMOTE), and feature importance analysis provides an effective, robust, and interpretable framework for predicting IMDb movie rating categories under subjective and imbalanced data conditions. From a practical perspective, these findings offer valuable implications for digital platforms such as IMDb and streaming services by enabling more accurate and stable rating predictions, which can enhance personalized recommendation systems, content discovery mechanisms, and user engagement strategies, as well as support data-driven decision-making in content acquisition, promotion, and distribution planning. The novelty of this research lies in the systematic integration of ensemble learning, imbalance-aware modeling, and explainable machine learning into a unified analytical framework, establishing a theoretically grounded and practically applicable baseline model that extends beyond purely performance-oriented approaches in previous studies. In the context of recommendation system development and film industry analytics, the proposed Random Forest model contributes by delivering reliable classification performance while maintaining transparency through feature importance analysis, thereby facilitating explainable and trustworthy intelligent decision-support systems. Furthermore, the most relevant direction for future research involves incorporating unstructured data sources, particularly textual user reviews and sentiment-based features, alongside structured movie

attributes, using advanced natural language processing techniques to further enhance predictive accuracy, contextual understanding, and interpretability of movie rating prediction models.

Reference

Ahmad, A., & Khalid, H. (2021). Movie rating prediction using machine learning techniques. *Journal of Information Science*, *47*(4), 555–566. https://doi.org/10.1177/0165551520942213

Breiman, L. (2001). Machine Learning, Volume 45, Number 1 - SpringerLink. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

He, H., & Ma, Y. (2021). Imbalanced learning foundations and algorithms. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, *11*(4). https://doi.org/10.1002/widm.1400

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2 ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Liu, Y., Zhang, X., & Wang, J. (2022). Predicting movie popularity and ratings using machine learning approaches. *Expert Systems with Applications*, *198*, 116742. https://doi.org/10.1016/j.eswa.2022.116742

Lundberg, S. M., & Lee, S.-I. (2020). A unified approach to interpreting model predictions. *Nature Machine Intelligence*, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9

Martínez, A., & Martínez, J. (2021). Analysis of IMDb dataset for movie success prediction. *Procedia Computer Science*, *179*, 547–554. https://doi.org/10.1016/j.procs.2021.01.036

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Powers, D. M. W. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. https://arxiv.org/abs/2010.16061

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sarker, I. H. (2021). Machine learning algorithms and applications. *SN Computer Science*, *2*(3). https://doi.org/10.1007/s42979-021-00592-x

Song, J., & Kim, Y. (2020). Movie rating prediction using ensemble learning. *Applied Sciences*, *10*(18), 6352. https://doi.org/10.3390/app10186352

Wang, Z., Li, Y., & Liu, H. (2023). Explainable random forest for decision support systems. *Expert Systems with Applications*, *213*, 118882. https://doi.org/10.1016/j.eswa.2022.118882

Wu, J., Guo, Y., Zhou, H., Shen, L., & Liu, L. (2020). Vehicular Delay Tolerant Network Routing Algorithm Based on Bayesian Network. *IEEE Access*, *8*, 18727–18740. https://doi.org/10.1109/ACCESS.2020.2967898

Yang, L., Qian, Y., & Wang, Y. (2020). Movie rating prediction via ensemble classifiers. *IEEE Access*, *8*, 123210–123219. https://doi.org/10.1109/ACCESS.2020.3008024

Zhou, Z.-H. (2021). Ensemble learning. In C. Sammut & G. I. Webb (Ed.), *Encyclopedia of Machine Learning and Data Mining* (hal. 1–9). Springer. https://doi.org/10.1007/978-1-4899-7687-1_453