Perbanas International Conference on Economics, Business, Management, Accounting and IT (PROFICIENT) 2026
"Integrity-Driven Innovation: Reimagining Business, Finance, and Digital Transformation for Sustainable Impact"

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Predicting Book Popularity on Goodreads: A Comparative Study of Linear Regression, Decision Tree, and Random Forest

Mohammad Harits Akbar[1,a)], Widya Ayu Purwati [2,b)], Elliana Gautama[3,c)]

[1,2,3]*Data Science Study Program, Faculty of Information Technology, Perbanas Institute Jakarta*

[a)]Corresponding author: mohammad.harits01@perbanas.id,
[b)]widya.ayu09@perbanas.id,
[c)]elliana@perbanas.id

Abstract. *The increasing volume of digital book data on social reading platforms such as Goodreads has created an opportunity to analyze factors that influence book popularity using machine learning methods. This study aims to identify which features significantly affect book popularity scores and to compare the predictive performance of three machine learning models: Linear Regression, Decision Tree, and Random Forest. The dataset used consists of 1,296 book records from the Goodreads favorite book list, comprising variables such as has_series, series_number, avg_rating, total_rating, and voters. After preprocessing including the removal of missing values and duplicate data—1,191 records were retained for modeling. The models were evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). Results indicate that the Random Forest model achieved the best performance with R^2 = 0.9722, MAE = 15.82, and RMSE = 88.13, outperforming both Decision Tree (R^2 = 0.9676) and Linear Regression (R^2 = 0.8479). Feature importance analysis revealed that the voters variable has the most dominant influence on book popularity with an importance score of 0.9794. This study contributes to the understanding of digital reading behavior and provides a foundation for developing more accurate book recommendation systems.*

Keywords: Book Popularity, Decision Tree, Linear Regression, Machine Learning, Random Forest

INTRODUCTION

The rapid development of digital technology has transformed how readers access, evaluate, and share their reading experiences. Social reading platforms such as Goodreads have become rich repositories of user-generated data, enabling researchers to analyze patterns of book popularity at scale. With millions of books rated and reviewed by users worldwide, understanding which factors drive popularity has significant implications for publishers, authors, and recommendation system developers (Maity et al., 2018).

Popularity in the context of social reading platforms is typically operationalized through composite scores derived from user engagement metrics, including the number of votes, average ratings, and total number of ratings received by a book. Prior research has demonstrated that user engagement features—rather than purely content-based features tend to be stronger predictors of popularity (Wijaya et al., 2023); (Maity et al., 2018). However, a systematic comparison of machine learning methods for predicting book popularity scores using structured tabular data from Goodreads remains underexplored.

This study uses a favorite books dataset obtained from the Kaggle platform. The variables used in this study include has_series, series_number, avg_rating, total_rating, voters, and score as indicators of book popularity. This research applies three machine learning methods, namely Linear Regression, Decision Tree Regression, and Random Forest Regression, to analyze the factors influencing book popularity. In addition,

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

model performance evaluation is conducted using evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score to compare the prediction accuracy of each method

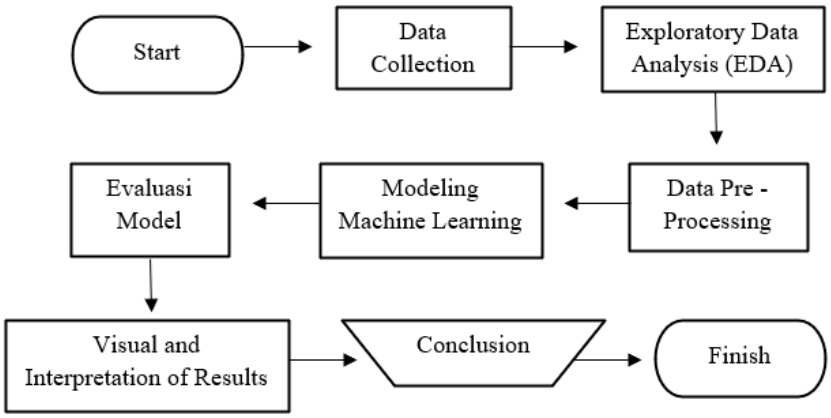

FIGURE 1. stage in research

Machine learning regression methods offer a powerful means to model the relationship between multiple input features and a continuous target variable such as a popularity score. Linear Regression provides an interpretable baseline by assuming a linear relationship between predictors and the outcome. Decision Tree Regression captures non-linear patterns by partitioning the feature space hierarchically (Schonlau & Zou, 2020). Random Forest, as an ensemble of decision trees, further reduces variance and improves generalization by averaging predictions across multiple trees (Breiman, 2001)

The research question guiding this study is: which machine learning method produces the most accurate predictions of book popularity scores, and which features contribute most significantly to those predictions? To answer this, the study uses a dataset of 1,296 books scraped from the Goodreads platform and compares the performance of Linear Regression, Decision Tree, and Random Forest Regression models. The results are evaluated using standard regression metrics: MAE, RMSE, and R^2.

The contributions of this study are threefold: (1) it provides an empirical comparison of three widely-used machine learning models for predicting book popularity; (2) it identifies voters as the dominant feature influencing book popularity scores; and (3) it offers practical insights for the development of book recommendation systems based on user engagement data.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

LITERATURE REVIEW

Book Popularity Prediction

Research on predicting book popularity has grown alongside the availability of large-scale social reading datasets. (Maity et al., 2018) investigated book popularity on Goodreads by analyzing user engagement and author prestige features, achieving a correlation coefficient of approximately 0.61 in predicting vote counts. Their work established that user engagement metrics such as average rating, number of ratings, and user-generated shelving behavior are key determinants of book popularity. (Wijaya et al., 2023) extended this line of inquiry by developing a prediction model for book popularity using data from Goodreads' 'To Read' and 'Worst' book lists, demonstrating the applicability of multiple machine learning methods for popularity modeling. (Verma et al., 2023) further explored rating prediction from Goodreads book reviews, highlighting the role of textual features alongside structured metadata.

Linear Regression

Linear Regression is one of the most fundamental supervised learning algorithms for regression tasks. It models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. While its interpretability makes it valuable for initial analysis, Linear Regression assumes a linear relationship between predictors and the target variable, which can limit its performance on datasets with complex, non-linear patterns (G. James et al., 2023). In comparative studies, Linear Regression typically serves as a baseline against which more sophisticated models are evaluated (Sayed et al., 2025).

Decision Tree Regression

Decision Tree Regression is a non-parametric supervised learning method that partitions the feature space into hierarchical decision rules to predict a continuous target variable. Each internal node represents a decision based on a feature value, and each leaf node represents a predicted output—typically the mean of the target values in that subset (Schonlau & Zou, 2020). Decision trees are well-suited for capturing non-linear relationships and interactions between variables. They also provide inherent feature importance measures, which facilitate the interpretation of which variables drive predictions. However, individual decision trees are prone to overfitting when grown to full depth (G. Chaudhary et al., 2023)

Random Forest Regression

Random Forest, introduced by(Breiman, 2001), is an ensemble method that constructs multiple decision trees on bootstrap samples of the training data and aggregates their predictions through averaging. This bagging approach significantly reduces variance compared to a single decision tree, improving generalization performance on unseen data. (Schonlau & Zou, 2020) demonstrated that Random Forest consistently outperforms Linear Regression on prediction tasks due to its ability to adapt to non-linear patterns. Multiple comparative studies have confirmed that Random Forest achieves superior performance over linear models and single decision trees onc complex structured datasets, producing lower MAE and RMSE values alongside higher R^2 scores (Sayed et al., 2025)

Feature Importance Analysis

Feature importance analysis is a key advantage of tree-based models such as Decision Tree and Random Forest. In Random Forest, feature importance is computed as the mean decrease in node impurity across all trees attributable to splits on each feature. Because Random Forest averages importance scores across many diverse trees, its estimates are considerably more stable than those from a single decision tree (Pallathadka et al., 2023). Research on social platform data has consistently found that user interaction metrics such as vote counts and engagement frequency exhibit the strongest predictive power for content popularity (Peters et al., 24).

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

METHODS

The dataset presented is a raw dataset containing 1,296 research data obtained from the Goodreads favorite book list, and each record includes the following variables: title_book (book title), title_series_book (series title), author_name (author name), has_series (binary indicator: 1 if the book belongs to a series, 0 otherwise), series_number (position number in the series), avg_rating (average user rating on a scale of 0–5), total_rating (total number of ratings received), voter (number of users who voted for the book in the popularity ranking), and score (book popularity score, which serves as the target variable).

TABLE 1. Dataset

	Unnamed: 0	title_book	title_series_book	author_name	has_series	series_number	avg_rating	total_rating	voters	score
0	0	The Hunger Games	The Hunger Games	Suzanne Collins	1.0	1.0	4.33	6086305.0	196.0	9286.0
1	1	Catching Fire	The Hunger Games	Suzanne Collins	1.0	2.0	4.29	2360261.0	138.0	3376.0
2	2	Mockingjay	The Hunger Games	Suzanne Collins	1.0	3.0	4.04	2217517.0	99.0	9442.0
3	3	Divergent	Divergent	Veronica Roth	1.0	1.0	4.20	2767407.0	97.0	9119.0
4	4	Vampire Academy	Vampire Academy	Richelle Mead	1.0	1.0	4.13	7029.0	86.0	8292.0
...	...	...	...	...	...	...	...	...	...	...
1290	1294	Dark Memories	The Phantom Diaries	Kailin Gow	1.0	2.0	4.18	0.0	1.0	0.0
1291	1295	Lament: The Faerie Queen's Deception	Books of Faerie	Maggie Stiefvater	1.0	1.0	3.67	1207.0	2.0	0.0
1292	1296	Parrotfish	NaN	Ellen Wittlinger	0.0	0.0	3.71	5671.0	1.0	0.0
1293	1297	Travels with Gannon and Wyatt: Great Bear Rain...	NaN	Patti Wheeler	0.0	0.0	4.19	0.0	1.0	0.0
1294	1298	United We Spy	Gallagher Girls	Ally Carter	1.0	6.0	4.44	9093.0	1.0	0.0

1295 rows × 10 columns

At this stage, the data has not yet gone through the cleaning or transformation process, so there are still several problems in the dataset. It can be seen that the raw data still contains missing values (NaN), especially in variables related to book series information such as title_series_book and series_number

Data Preprocessing

Prior to modeling, the dataset underwent several preprocessing steps. First, the irrelevant index column (Unnamed: 0) was removed. Second, non-numeric identifier variables (title_book, title_series_book, author_name) were excluded from the feature set. Third, records with missing values were removed using listwise deletion. Fourth, duplicate records were identified and eliminated.

TABLE 2. dataset variables after preprocessing stage

	title_book	has_series	series_number	avg_rating	total_rating	voters	score
0	The Hunger Games	1.0	1.0	4.33	6086305.0	196.0	9286.0
1	Catching Fire	1.0	2.0	4.29	2360261.0	138.0	3376.0
2	Mockingjay	1.0	3.0	4.04	2217517.0	99.0	9442.0
3	Divergent	1.0	1.0	4.20	2767407.0	97.0	9119.0
4	Vampire Academy	1.0	1.0	4.13	7029.0	86.0	8292.0

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

After preprocessing, 1,191 complete records were retained for analysis. The final feature set comprised five predictor variables: has_series, series_number, avg_rating, total_rating, and voters. The target variable was score.

Correlation Analysis

Correlation analysis was conducted to assess the bivariate relationships between features and the target variable. The voters variable exhibited the strongest correlation with score (r = 0.92), followed by total_rating (r = 0.53). The remaining variables avg_rating (r = 0.06), series_number (r = 0.14), and has_series (r = 0.12) showed relatively weak correlations with the popularity score.

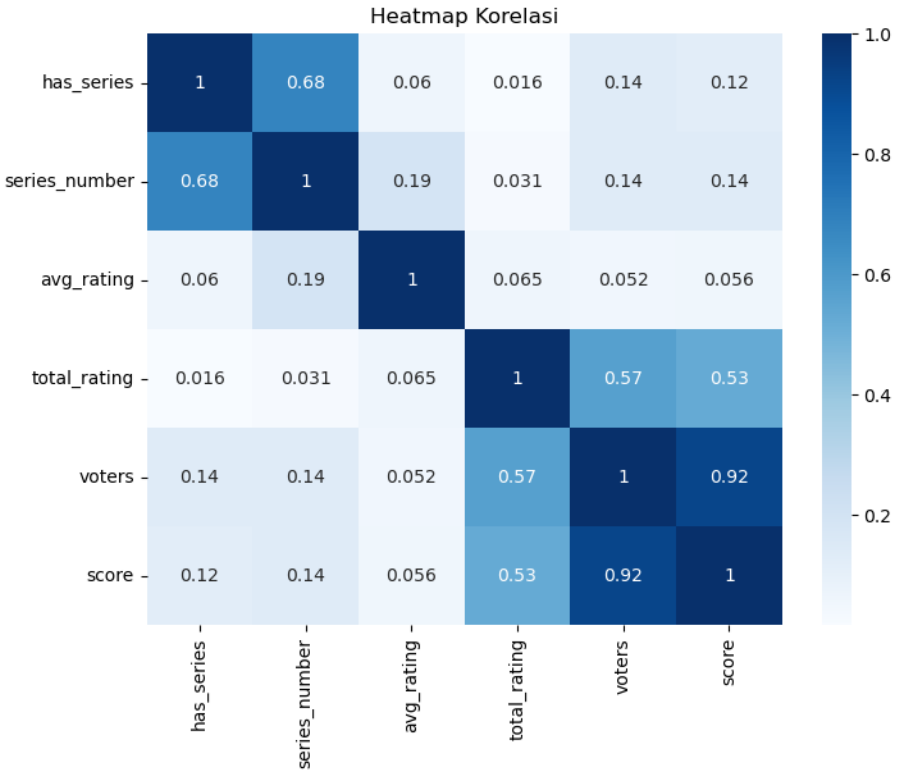

FIGURE 2. Correlation Heatmap of All Variables

Model Development

The dataset was split into training and testing sets using an 80:20 ratio, resulting in 952 records for training and 239 records for testing. A fixed random seed (random_state = 42) was used to ensure reproducibility. Three regression models were developed: (1) Linear Regression using Ordinary Least Squares; (2) Decision Tree Regression using the CART algorithm with default hyperparameters; and (3) Random Forest Regression using an ensemble of 100 decision trees (n_estimators = 100). All models were implemented using the Scikit-learn library in Python.

Model Evaluation

Model performance was evaluated using three regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). Higher R^2 values (closer to 1.0) and lower MAE and RMSE values indicate better model performance (G. James et al., 2023). The formulas for each metric are: $MAE = (1/n)\Sigma|y_i - \hat{y}$

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

$_i|$; RMSE = $\sqrt{[(1/n)\Sigma(y_i - \hat{y}_i)^2]}$; $R^2 = 1 - [\Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2]$, where y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of actual values

RESULTS AND DISCUSSION

Descriptive Statistics

Descriptive analysis revealed considerable variation in book popularity scores. The score variable ranged from 0 to 9,442 with a mean of 365.14 and a standard deviation of 1,208.59, indicating a highly right-skewed distribution. The voters variable similarly exhibited high variance (range: 1–196, mean: 6.26), consistent with power-law distributions commonly observed on social platforms. These descriptive patterns suggest that popularity is concentrated among a small number of books, while the majority receive minimal engagement.

Model Performance Comparison

Table 1 presents the performance metrics for all three models on the test set (239 records).

TABLE 3. Performance Comparison of Regression Models

Model	MAE	RMSE	R^2 Score
Linear Regression	112.80	206.23	0.8479
Decision Tree	14.76	95.11	0.9676
Random Forest	15.82	88.13	0.9722

Source: Primary data processed by authors (2025)

The Linear Regression model achieved an R^2 of 0.8479, indicating that the five predictor variables explain approximately 84.79% of the variance in popularity scores. The relatively high MAE (112.80) and RMSE (206.23) suggest the model struggles to accurately predict books with extreme popularity scores, consistent with the non-linear relationship between predictors and the target variable.

The Decision Tree Regression model demonstrated a substantial improvement, achieving an R^2 of 0.9676, MAE of 14.76, and RMSE of 95.11. The dramatic reduction in error metrics (MAE reduced by approximately 87% compared to Linear Regression) confirms that the popularity-feature relationship involves significant non-linear patterns that the tree-based partitioning approach is better equipped to model. These findings are consistent with (Schonlau & Zou, 2020) and (G. Chaudhary et al., 2023)

The Random Forest Regression model achieved the best overall performance with an R^2 of 0.9722, MAE of 15.82, and RMSE of 88.13. Although the MAE is marginally higher than Decision Tree (15.82 vs. 14.76), the lower RMSE (88.13 vs. 95.11) indicates that Random Forest generates fewer large prediction errors. This reflects the ensemble nature of Random Forest: by averaging predictions across 100 decision trees, it reduces the variance responsible for large individual tree errors (Breiman, 2001; (Sayed et al., 2025).

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

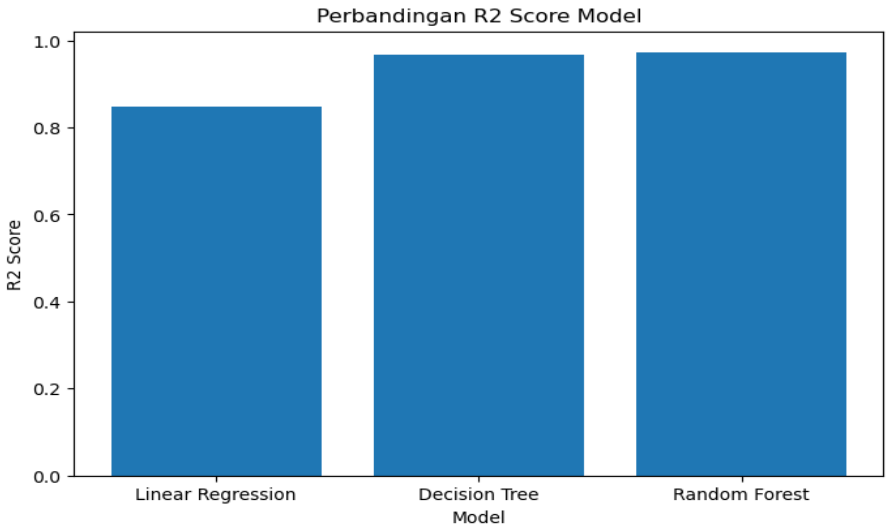

FIGURE 3. Comparison of R^2 Score Across Three Regression Models

Feature Importance Analysis

Table 4 presents the feature importance scores derived from the trained Random Forest model.

TABLE 4. Feature Importance Scores from Random Forest Model

Rank	Feature	Importance Score
1	voters	0.9794
2	total_rating	0.0093
3	avg_rating	0.0063
4	series_number	0.0044
5	has_series	0.0006

Source: Primary data processed by authors (2025)

Feature importance analysis reveals that voters dominates all other predictors with an importance score of 0.9794 meaning the number of user votes accounts for approximately 97.94% of the predictive power in the model. This finding is strongly corroborated by the correlation analysis (r = 0.92) and is consistent with (Maity et al., 2018), who identified user engagement as the primary driver of Goodreads book popularity. The remaining features total_rating (0.0093), avg_rating (0.0063), series_number (0.0044), and has_series (0.0006) contribute minimally, suggesting that structural characteristics of books have limited direct influence on popularity relative to direct user interaction.

The dominance of voters over avg_rating is a theoretically important finding: it implies that the volume of user engagement matters far more than its quality. This aligns with broader research on online platforms demonstrating that interaction volume is a stronger predictor of content popularity than quality metrics (Peters et al., 24); (Pallathadka et al., 2023)

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

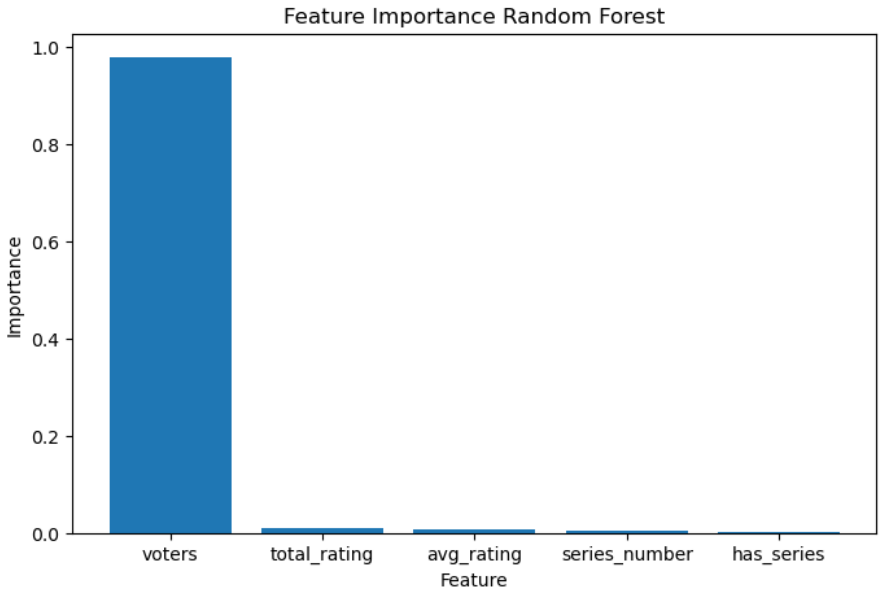

FIGURE 4. Feature Importance Scores of Random Forest Model

Top 10 Most Popular Books

Table 3 presents the top 10 books in the dataset ranked by their popularity score.

TABLE 5. Top 10 Most Popular Books by Score

Rank	Book Title	Score
1	Mockingjay	9,442
2	The Hunger Games	9,286
3	Divergent	9,119
4	Vampire Academy	8,292
5	Harry Potter and the Deathly Hallows	8,066
6	Clockwork Angel	7,374
7	Harry Potter and the Goblet of Fire	6,674
8	Hush, Hush	6,382
9	Harry Potter and the Half-Blood Prince	6,343
10	Harry Potter and the Order of the Phoenix	5,336

Source: Primary data processed by authors (2025)

The top 10 most popular books are dominated by major young adult fiction series, including The Hunger Games trilogy, Harry Potter series, and other prominent franchises. This pattern suggests a strong community effect: books belonging to well-known series generate large readership communities, which in turn drives higher voter counts and consequently higher popularity scores. This indirect pathway explains why series books dominate the top rankings despite has_series having a low direct feature importance score.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

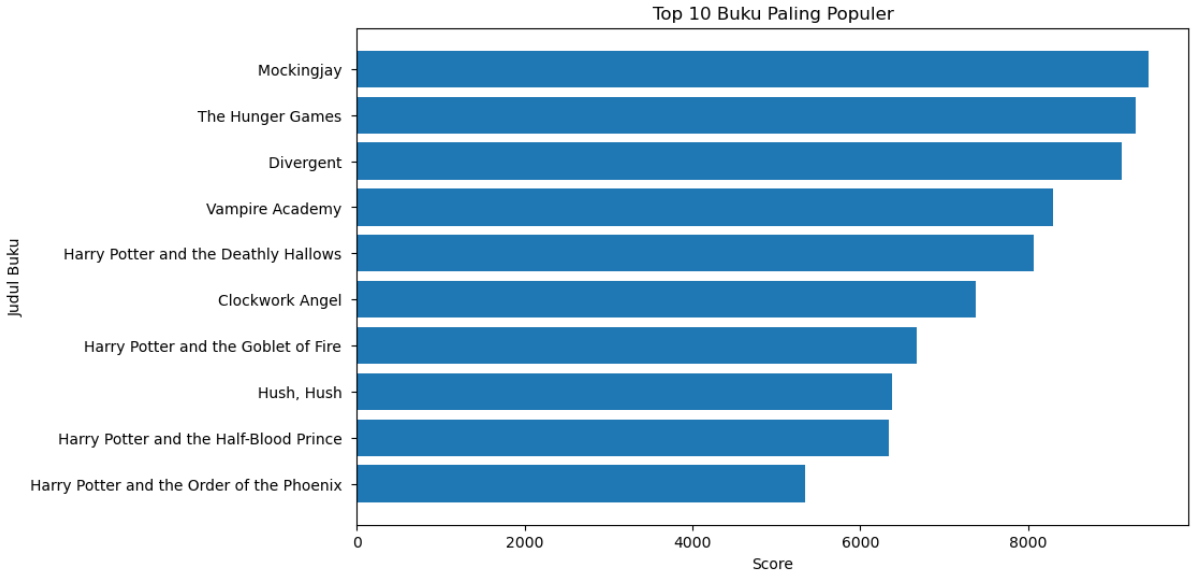

FIGURE 5. Bar Chart of Top 10 Most Popular Books by Score

CONCLUSIONS

This study compared three machine learning regression models for predicting book popularity scores based on structured metadata from the Goodreads platform. The following conclusions are drawn:

(1) The Random Forest model achieved the highest predictive performance (R^2 = 0.9722, RMSE = 88.13), followed by Decision Tree (R^2 = 0.9676, RMSE = 95.11) and Linear Regression (R^2 = 0.8479, RMSE = 206.23). The superior performance of tree-based methods confirms that the relationship between book features and popularity scores is fundamentally non-linear.

(2) The voters variable is overwhelmingly the most important predictor of book popularity, with an importance score of 0.9794 accounting for approximately 97.94% of the Random Forest model's predictive power.

(3) Structural book characteristics such as has_series and series_number have minimal direct predictive power, while avg_rating also contributes marginally compared to direct user engagement metrics.

Implications

These findings have practical implications for book recommendation system development. Recommendation algorithms that incorporate voter engagement data as a primary feature are likely to produce more accurate popularity predictions than those relying solely on average ratings. For publishers and authors, the results suggest that strategies aimed at increasing user interaction and community engagement rather than focusing exclusively on rating improvement are more effective for elevating book visibility on social reading platforms.

Limitations and Future Research

This study has several limitations. The dataset is limited to books featured in the Goodreads favorite book list, which may not represent the broader catalog. The study relies on structured tabular features and does not incorporate textual features from reviews or synopses. Future research could address these limitations by using larger datasets, incorporating natural language processing of review text, and

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

experimenting with advanced methods such as Gradient Boosting or XGBoost. Longitudinal analyses examining how popularity evolves over time would also contribute valuable insights.

Acknowledgments

The authors would like to thank all parties who contributed to the completion of this research.

References

Breiman, L. (2001). *Random forests. Machine Learning*. 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Chaudhary, G., Gavhane, P. R., Bisen, D., & Arjaria, S. K. (2023). *Customer purchasing behavior prediction using machine learning classification techniques. Journal of Ambient Intelligence and Humanized Computing*. 16133–16157.

James, G., Hastie, T., Witten, D., Tibshirani, R., Gundecha, A., Salhotra, S., & Lee, E. (2023). *An Introduction to Statistical Learning with Applications in Python. Springer.* .

Maity, S. K., Kumar, S., Bansal, M., & Mukherjee, A. (2018). *Understanding book popularity on Goodreads.*

Pallathadka, H., Wenda, A., Ramirez-Asis, E. H., & Asís-López, M. E. (2023). *Classification and prediction in student performance analysis using machine learning algorithms*. 3782–3785.

Peters, H., Bayer, J. B., Matz, S. C., Chi, Y., Vaid, S. S., & Harari, G. M. (24 C.E.). *Context-aware prediction of active and passive user engagement: Evidence from a large online social platform.* .

Sayed, A. F. A., Arafa, M. A., El-Nimr, N. A., Banawan, K. A. S., & Abdou, M. S. (2025). *Comparison of machine learning classification and regression models for prediction of academic performance among postgraduate public health students*. 1–15.

Schonlau, M., & Zou, R. Y. (2020). *The random forest algorithm for statistical learning*. 3–29.

Verma, A., Baliyan, N., Gera, P., & Jindal, S. (2023). *Predicting corresponding ratings from Goodreads book reviews*. 537.

Wijaya, R. A., Staniswinata, S., Clarin, M., Qomariyah, N. N., & Manuaba, I. B. K. (2023). *Prediction model of book popularity from Goodreads "To Read" and "Worst" books*. 1–7.