JURNAL ILMIAH PERPUSTAKAAN, SISTEM INFORMASI, INFORMATIKA & ELEKTRO
Volume 1, Nomor 1, Juni 2026
. hlm. 56 -72

Analisis Faktor yang Mempengaruhi Rating Buku Menggunakan Teknik Data Mining

Maryam Risqa Andiani[1], Egi Abinowi[2]
[1, 2]Fakultas Ilmu Sosial dan Ilmu Politik, Universitas Widyatama, Jl. Cikutra no 204 A Bandung Jawa Barat, Indonesia 40124
E-mail korespondensi: [1]maryam.andiani@widyatama.ac.id, [2]egi.abinowi@widyatama.ac.id,

Keywords: book ratings, data mining, goodreads, library, metadata

Abstract
The development of information technology has transformed how people interact with books, particularly through rating and reviewing on digital platforms, which generates large-scale data. Although the utilization of data mining in libraries continues to expand, research comprehensively analyzing the impact of metadata characteristics on average book ratings remains limited. This study aims to analyze the bibliographical metadata factors that influence the average rating of books using the Goodreads dataset. The method employed is a quantitative approach using data mining techniques with the random forest regression algorithm on ten thousand book records. The results indicate that the relationships between variables are non-linear, with the work text reviews count and total work ratings count being the most influential factors in predicting the average rating, while the original publication year has the lowest impact. This study concludes that reader appreciation in the digital era is heavily driven by the intensity of community interaction and discussion surrounding the book. These findings are vital as an empirical foundation for librarians in implementing data-driven decision-making for collection development and optimizing library recommendation services.

Kata kunci: data mining, goodreads, metadata, perpustakaan, rating buku

Abstrak
Perkembangan teknologi informasi mengubah cara masyarakat berinteraksi dengan buku, salah satunya melalui pemberian penilaian dan ulasan di platform digital yang menghasilkan data berskala besar. Meskipun pemanfaatan penambangan data di perpustakaan terus berkembang, penelitian yang menganalisis faktor karakteristik metadata terhadap nilai rata-rata rating buku secara komprehensif masih terbatas. Penelitian ini bertujuan untuk menganalisis faktor-faktor metadata bibliografis yang memengaruhi nilai rata-rata rating buku pada dataset Goodreads. Metode yang digunakan adalah pendekatan kuantitatif dengan teknik penambangan data menggunakan algoritma regresi *random forest* terhadap sepuluh ribu data buku. Hasil penelitian menunjukkan bahwa hubungan antarvariabel bersifat non-linear, dengan jumlah ulasan teks dan jumlah total rating menjadi faktor yang paling berpengaruh dalam memprediksi

nilai rata-rata rating buku, sementara tahun publikasi memiliki pengaruh terendah. Penelitian ini menyimpulkan bahwa apresiasi pembaca di era digital sangat dipengaruhi oleh intensitas interaksi dan diskusi komunitas di sekitar buku tersebut. Hasil ini penting sebagai dasar empiris bagi pustakawan dalam menerapkan pengambilan keputusan berbasis data untuk pengembangan koleksi dan optimalisasi layanan rekomendasi perpustakaan.

PENDAHULUAN

Perkembangan teknologi informasi telah mengubah cara masyarakat dalam memperoleh, mengelola, dan memanfaatkan informasi. Salah satu perubahan yang paling signifikan ditunjukkan oleh meningkatnya penggunaan platform digital sebagai media untuk mencari, membaca, serta berbagi informasi mengenai buku. *Goodreads* merupakan salah satu platform yang banyak digunakan oleh pembaca di seluruh dunia untuk menemukan koleksi buku, memberikan penilaian (*rating*), serta menulis ulasan (*review*). Aktivitas tersebut menghasilkan data dalam jumlah yang sangat besar dan terus bertambah setiap hari sehingga menjadi salah satu bentuk implementasi *big data* dalam bidang informasi. Melalui proses analisis data yang sistematis, *Big Data* memungkinkan organisasi untuk mengidentifikasi pola, tren, dan hubungan tersembunyi yang sulit diperoleh melalui pendekatan konvensional (Nurina et al., 2024). Dalam bidang Perpustakaan dan Sains Informasi, Big Data dimanfaatkan untuk menganalisis perilaku pengguna, pola pemanfaatan koleksi, serta mendukung pengambilan keputusan berbasis data dalam pengelolaan layanan informasi dan perpustakaan (Dotulong et al., 2026).

Karakteristik Big Data umumnya dijelaskan melalui kerangka lima V (5V) yaitu *volume*, *velocity*, *variety*, *veracity*, dan *value* di mana pengelolaannya memerlukan dukungan teknologi, arsitektur, serta perangkat analitik yang dirancang khusus agar mampu mengolah data dalam jumlah besar secara efisien dan menghasilkan nilai tambah (*value*) (Suryantari et al., 2026). *Data mining* merupakan proses pencarian dan penemuan pengetahuan (*knowledge discovery*) dari sekumpulan data bervolume sangat besar melalui identifikasi pola dan hubungan yang valid menggunakan teknik analisis statistik, *machine learning*, serta visualisasi informasi (Firdaus, 2017). Menurut Baskara et al. (2025), tidak terbatas pada pengategorian dan prediksi nilai saja, fungsi *data mining* juga mencakup identifikasi hubungan serta tren antar-item dalam dataset berskala besar demi mewujudkan proses pengambilan keputusan organisasi yang berbasis bukti empiris. Dalam konteks Perpustakaan dan Sains Informasi, Mengadopsi pemanfaatan teknologi analitik yang telah berhasil diimplementasikan pada evaluasi mutu pelayanan berbasis data digital, penerapan *data mining* dalam dunia perpustakaan kini telah berkembang pada analisis perilaku pemustaka, evaluasi penggunaan koleksi, pengembangan sistem rekomendasi, hingga

pengelolaan layanan informasi berbasis bukti (*evidence-based librarianship*) (Faradillah et al., 2025).

Salah satu sumber data yang banyak dimanfaatkan dalam penelitian adalah *Goodreads Books Dataset* yang tersedia di Kaggle memuat sekitar 10.000 data buku dan 23 atribut metadata, sehingga sesuai digunakan sebagai objek penelitian *data mining*. Metadata tersebut merepresentasikan karakteristik sebuah buku sekaligus preferensi pembaca terhadap koleksi yang mereka baca. Khadijah et al. (2025) menyatakan bahwa *Goodreads* sebagai salah satu jejaring sosial berbasis buku menawarkan ruang interaksi baru bagi para pembaca untuk saling berbagi pendapat, memberikan penilaian, dan mendiskusikan buku secara terbuka, sehingga hal ini menjadikannya sebagai sumber data yang relevan untuk mengkaji resepsi pembaca secara langsung. Worrall (2019) juga menjelaskan bahwa *Goodreads* tidak hanya berfungsi sebagai media sosial bagi pembaca, tetapi telah berkembang menjadi komunitas informasi yang mampu menggambarkan interaksi pengguna dengan sumber informasi. Dengan demikian, metadata yang tersedia pada *Goodreads* memiliki potensi besar untuk dimanfaatkan sebagai sumber analisis dalam pengembangan layanan perpustakaan berbasis data.

Pemanfaatan data *Goodreads* dalam penelitian telah berkembang cukup pesat. Kusumaningtyas & Ramdhani (2025) menganalisis tren dan perilaku literasi masyarakat dengan membedah data sekunder berupa *rating* serta ulasan buku yang ditinggalkan pengguna di *Goodreads*. Menurut Tupan (2024) penerapan *machine learning* dan analisis data di perpustakaan dapat membantu pustakawan mengoptimalkan proses pengembangan koleksi serta memfasilitasi tingkat manajerial dalam pengambilan keputusan. Sementara itu, Chauhan (2021, dalam Galuh et al., 2025) membuktikan bahwa integrasi algoritma *data mining* dengan teknik *filtering* mampu meningkatkan akurasi sistem rekomendasi. Penelitian serupa oleh Listianto et al. (2024) mengeksplorasi potensi *data mining*, terutama menggunakan metode *association rule*, dalam meningkatkan efisiensi penempatan buku di perpustakaan, Sementara itu, Indah et al. (2026) melalui kajian literatur sistematis menunjukkan bahwa riset mengenai sistem rekomendasi buku di perpustakaan digital sejauh ini masih didominasi oleh dua pendekatan utama, yaitu *collaborative filtering* dan *content-based filtering* . Hasil-hasil penelitian tersebut menunjukkan bahwa penerapan *data mining* pada data buku digital masih didominasi oleh pengembangan sistem rekomendasi, analisis sentimen, dan klasifikasi koleksi.

Meskipun demikian, penelitian yang secara khusus menganalisis faktor-faktor yang memengaruhi nilai rata-rata *rating* buku berdasarkan karakteristik metadata masih relatif terbatas. Berdasarkan penelitian Kusumaningtyas & Ramdhani (2025), Tupan (2024), Chauhan (2021, dalam Galuh et al., 2025), Listianto et al. (2024), serta Indah et al. (2026), dapat diketahui bahwa pemanfaatan data *Goodreads* dan teknik *data mining* di perpustakaan masih didominasi oleh

pengembangan sistem rekomendasi, analisis sentimen, penempatan, dan pengembangan koleksi. Penelitian-penelitian tersebut belum mengkaji secara komprehensif hubungan antaratribut metadata buku seperti jumlah halaman (*num_pages*), jumlah pemberi *rating* (*ratings_count*), jumlah ulasan (*work_text_reviews_count*), tahun publikasi (*original_publication_year*), dan jumlah edisi (*books_count*) terhadap nilai *average rating.* Padahal hubungan antarvariabel tersebut berpotensi memberikan informasi mengenai karakteristik buku yang memperoleh apresiasi tinggi dari pembaca. Kondisi ini menunjukkan adanya kesenjangan penelitian (*research gap*) yang perlu diisi melalui analisis metadata menggunakan teknik data mining.

Urgensi penelitian ini terletak pada meningkatnya kebutuhan perpustakaan untuk menerapkan pengambilan keputusan berbasis data (*data-driven decision making*). Informasi mengenai karakteristik buku yang memperoleh rating tinggi dapat dimanfaatkan sebagai dasar dalam pengembangan koleksi, evaluasi kualitas koleksi, serta penyusunan layanan rekomendasi yang lebih sesuai dengan kebutuhan pengguna. Oleh karena itu, analisis terhadap faktor-faktor yang memengaruhi rating buku menjadi penting untuk mendukung transformasi digital perpustakaan.

Untuk menjawab kesenjangan tersebut, penelitian ini menggunakan *Goodreads Books Dataset* yang berisi metadata buku dalam jumlah besar sehingga memungkinkan eksplorasi hubungan antarvariabel menggunakan teknik *data mining*. Pendekatan ini diharapkan mampu menghasilkan informasi yang tidak hanya menjelaskan karakteristik buku dengan rating tinggi, tetapi juga memberikan dasar empiris bagi pengembangan koleksi dan layanan informasi berbasis data. Dengan demikian, penelitian ini berkontribusi terhadap pengembangan penerapan *big data* dan *data mining* dalam bidang Perpustakaan dan Sains Informasi.

Untuk menjembatani kesenjangan tersebut, penelitian ini menawarkan kebaruan (*novelty*) berupa eksplorasi faktor-faktor yang memengaruhi nilai rata-rata rating buku (*average_rating*) menggunakan *Goodreads Books Dataset* yang diperoleh dari Kaggle (file *books.csv*). Berbeda dengan penelitian sebelumnya yang didominasi oleh pengembangan sistem rekomendasi atau analisis sentimen, penelitian ini berfokus pada penerapan teknik data mining untuk mengidentifikasi faktor-faktor metadata bibliografis yang memengaruhi nilai rata-rata rating buku. Variabel yang dianalisis meliputi jumlah pemberi rating (ratings_count), jumlah akumulasi rating pada suatu karya (work_ratings_count), jumlah ulasan teks (work_text_reviews_count), jumlah edisi buku (books_count), dan tahun publikasi asli (original_publication_year). Melalui pendekatan Random Forest Regression, penelitian ini bertujuan mengungkap tingkat pengaruh masing-masing variabel terhadap average_rating sehingga dapat memberikan dasar empiris bagi pemanfaatan big data dalam pengembangan koleksi, evaluasi koleksi, dan layanan informasi pada bidang Perpustakaan dan Sains Informasi.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menganalisis faktor-faktor yang memengaruhi nilai rata-rata rating buku berdasarkan metadata pada *Goodreads Books Dataset* menggunakan teknik *data mining*. Hasil penelitian ini diharapkan dapat memberikan kontribusi teoretis terhadap pengembangan kajian *big data* dan *data mining* dalam bidang Perpustakaan dan Sains Informasi. Secara praktis, penelitian ini dapat menjadi landasan pengambilan keputusan bagi lembaga dokumentasi dalam pengembangan koleksi, evaluasi kualitas koleksi, serta penyusunan layanan rekomendasi buku berbasis data (*data-driven decision making*).

METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan analisis data sekunder menggunakan teknik *data mining*. Data penelitian berasal dari *Goodreads* Books Dataset yang diperoleh melalui platform Kaggle dalam format CSV (Bhushan, 2018). Dataset yang digunakan berupa file *books.csv* yang terdiri atas 10.000 data buku dan 23 atribut metadata.

1. Sumber Data dan Variabel Penelitian

Sumber data penelitian adalah *Goodreads Books Dataset* yang diperoleh dari Kaggle. Dataset yang digunakan berupa file books.csv yang terdiri atas 10.000 data buku dan 23 atribut (Bhushan, 2018) . Setiap observasi merepresentasikan satu judul buku yang dilengkapi metadata bibliografis dan data interaksi pengguna *Goodreads*. Subjek penelitian adalah seluruh data buku yang terdapat pada file books.csv, sedangkan objek penelitian adalah hubungan antara metadata buku dengan nilai rata-rata rating (average_rating).

2. Identifikasi Variabel Penelitian

Penelitian menggunakan satu variabel dependen dan lima variabel independen yang dipilih berdasarkan ketersediaan data pada dataset serta relevansinya dengan tujuan penelitian.

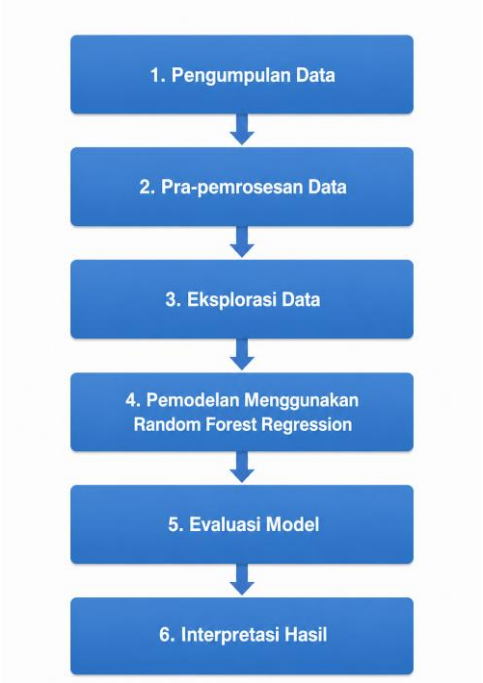
Tabel 1. Variabel Penelitian

Jenis Variabel	Nama Variabel	Keterangan
Dependen	average_rating	Nilai rata-rata rating buku
Independen	ratings_count	Jumlah pengguna yang memberikan rating
Independen	work_ratings_count	Jumlah total rating pada setiap karya
Independen	work_text_reviews count	Jumlah ulasan teks yang diberikan pengguna
Independen	books_count	Jumlah edisi buku yang diterbitkan
Independen	original_publication _year	Tahun publikasi pertama buku

Variabel tersebut dipilih karena secara konseptual mampu merepresentasikan tingkat popularitas, intensitas interaksi pembaca, serta karakteristik publikasi buku yang diduga memengaruhi nilai rata-rata *rating*.

3. Prosedur Penelitian

Penelitian dilaksanakan melalui beberapa tahapan yang disajikan pada Gambar 1.



Gambar 1. Tahapan Prosedur Penelitian

1. Pengumpulan Data

Dataset *Goodreads* Books diunduh dari platform Kaggle dalam format CSV. Selanjutnya dipilih file books.csv sebagai sumber data utama penelitian karena memuat metadata buku yang relevan dengan tujuan penelitian.

2. Pra-pemrosesan Data (*Data Preprocessing*)

Tahap *preprocessing* dilakukan untuk meningkatkan kualitas data sebelum proses analisis. Kegiatan yang dilakukan meliputi:

a. memeriksa struktur dan tipe data setiap atribut;
b. mengidentifikasi data duplikat;
c. memeriksa keberadaan *missing values* pada atribut penelitian;
d. memilih atribut yang digunakan dalam analisis
e. menghapus atribut yang tidak relevan, seperti ISBN, URL gambar, judul buku, dan identitas internal dataset; dan

f. memastikan seluruh variabel numerik berada pada format yang sesuai untuk proses pemodelan.

3. Eksplorasi Data (*Exploratory Data Analysis*)

Tahap eksplorasi dilakukan untuk memperoleh gambaran karakteristik dataset menggunakan analisis statistik deskriptif. Statistik yang dihitung meliputi nilai rata-rata (*mean*), median, simpangan baku (*standard deviation*), nilai minimum, dan maksimum dari setiap variabel numerik. Selain itu dilakukan visualisasi distribusi data dan analisis korelasi Pearson untuk mengetahui hubungan awal antarvariabel.

4. Pemodelan Menggunakan Random Forest Regression

Tahap pemodelan dilakukan menggunakan algoritma Random Forest Regression karena mampu memodelkan hubungan nonlinier antarvariabel dan menghasilkan nilai *feature importance* untuk mengidentifikasi variabel yang paling berpengaruh terhadap variabel target (Amin & Utami, 2025). Sebelum proses pelatihan model, dataset dibagi menjadi 80% data pelatihan (*training set*) dan 20% data pengujian (*testing set*).Variabel independen yang digunakan meliputi ratings_count, work_ratings_count, work_text_reviews_count, books_count, dan original_publication_year, sedangkan variabel dependen adalah average_rating. Random Forest Regression dipilih karena mampu mengolah data berukuran besar, memodelkan hubungan nonlinier antarvariabel, mengurangi risiko *overfitting* melalui mekanisme *ensemble learning*, serta menghasilkan nilai *feature importance* yang dapat digunakan untuk mengidentifikasi variabel yang paling berpengaruh terhadap nilai rata-rata *rating* buku.

5. Evaluasi Model

Kinerja model dievaluasi menggunakan tiga ukuran, yaitu:

a. Coefficient of Determination (R^2) untuk mengukur kemampuan model dalam menjelaskan variasi data;
b. Mean Absolute Error (MAE) untuk mengukur rata-rata kesalahan prediksi; dan
c. Root Mean Square Error (RMSE) untuk mengukur besarnya kesalahan prediksi secara keseluruhan.

6. Interpretasi Hasil

Tahap akhir dilakukan dengan menganalisis nilai feature importance yang dihasilkan oleh Random Forest Regression untuk menentukan variabel metadata yang paling berpengaruh terhadap nilai rata-rata *rating*. Hasil analisis kemudian diinterpretasikan dalam konteks pengembangan koleksi, evaluasi koleksi, serta penyusunan layanan rekomendasi berbasis data pada bidang Perpustakaan dan Sains Informasi.

4. Instrumen Penelitian

Instrumen penelitian berupa *Goodreads Books Dataset* (books.csv) yang terdiri atas 10.000 data buku dengan 23 atribut metadata. Proses pengolahan data dilakukan menggunakan bahasa pemrograman Python dengan pustaka Pandas untuk manipulasi data, NumPy untuk komputasi numerik, Matplotlib untuk visualisasi data, serta Scikit-learn untuk implementasi algoritma Random Forest Regression dan evaluasi model.

5. Teknik Analisis Data

Analisis data dilakukan dalam dua tahap. Tahap pertama menggunakan analisis statistik deskriptif dan korelasi Pearson untuk menggambarkan karakteristik data serta hubungan awal antarvariabel. Tahap kedua menggunakan teknik *data mining* melalui algoritma Random Forest Regression untuk mengidentifikasi faktor-faktor yang memengaruhi nilai rata-rata *rating* buku. Tingkat kepentingan setiap variabel ditentukan berdasarkan nilai *feature importance* yang dihasilkan model. Seluruh hasil analisis disajikan dalam bentuk tabel, grafik, dan interpretasi sehingga dapat digunakan sebagai dasar pengambilan keputusan berbasis data dalam pengembangan koleksi perpustakaan.

HASIL

1. Statistik Deskriptif

Statistik deskriptif dilakukan untuk memberikan gambaran karakteristik data sebelum proses pemodelan menggunakan teknik *data mining*. Dataset yang digunakan merupakan *Goodreads* Books Dataset yang terdiri atas 10.000 data buku dengan 23 atribut. Penelitian ini menggunakan lima variabel independen, yaitu ratings_count, work_ratings_count, work_text_reviews_count, books_count, dan original_publication_year, serta satu variabel dependen yaitu average_rating.

Tabel 2. Statistik Deskriptif Variabel Penelitian

Statistik	average_rating	ratings_count	work_ratings_count	work_text_reviews_count	books_count	original_publication_year
Count	10.000	10.000	10.000	10.000	10.000	9.979
Mean	4,002	54.001,24	59.687,32	2.919,96	75,71	1.981,99
Std. Deviasi	0,254	157.370,06	167.803,79	6.124,38	170,47	152,58
Minimum	2,470	2.716	5.510	3	1	-1750
Kuartil 1 (25%)	3,850	13.568,75	15.438,75	694	23	1990
Median (50%)	4,020	21.155,50	23.832,50	1.402	40	2004
Kuartil 3 (75%)	4,180	41.053,50	45.915,00	2.744,25	67	2011

Statistik	average_rating	ratings_count	work_ratings_count	work_text_reviews_count	books_count	original_publication_year
Maksimum	4,820	4.780.653	4.942.365	155.254	3.455	2017

Berdasarkan Tabel 2, diketahui bahwa variabel average_rating memiliki nilai rata-rata sebesar 4,002, dengan nilai minimum 2,470 dan maksimum 4,820. Hal ini menunjukkan bahwa sebagian besar buku pada *Goodreads* memperoleh penilaian yang relatif tinggi dari pengguna. Variabel ratings_count memiliki rata-rata 54.001 penilaian dengan nilai maksimum mencapai 4.780.653, sedangkan work_ratings_count memiliki rata-rata 59.687 dengan nilai maksimum 4.942.365. Kondisi tersebut menunjukkan adanya perbedaan tingkat popularitas yang cukup besar antar buku dalam dataset.

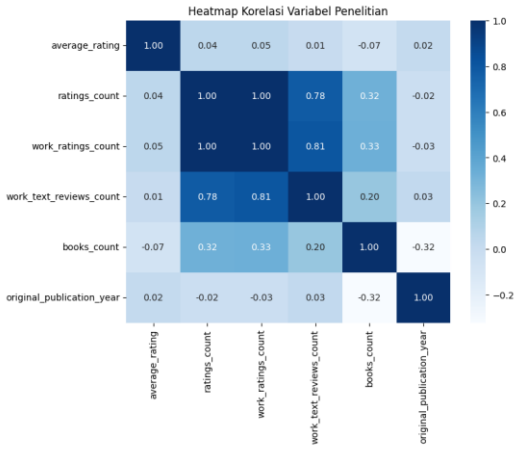
Variabel work_text_reviews_count memiliki rata-rata 2.919 ulasan teks dengan nilai maksimum 155.254, yang menunjukkan bahwa hanya sebagian buku memperoleh tingkat interaksi pengguna yang sangat tinggi melalui penulisan ulasan. Variabel books_count memiliki rata-rata 75 edisi dengan nilai maksimum mencapai 3.455 edisi, sedangkan original_publication_year memiliki rata-rata tahun publikasi 1981. Pada variabel original_publication_year terdapat 9.979 data valid, sehingga terdapat 21 data yang tidak memiliki informasi tahun publikasi. Secara keseluruhan, statistik deskriptif menunjukkan bahwa dataset memiliki variasi karakteristik yang cukup tinggi sehingga layak digunakan dalam proses pemodelan menggunakan teknik *data mining*.

2. Analisis Korelasi

Sebelum dilakukan pemodelan menggunakan Random Forest Regression, dilakukan analisis korelasi Pearson untuk melihat hubungan awal antarvariabel.

Tabel 3. Korelasi Variabel terhadap Average Rating

Variabel	Koefisien Korelasi
Ratings Count	**0.045**
Work Ratings Count	**0.045**
Work Text Reviews Count	**0.007**
Books Count	**−0.070**
Original Publication Year	**0.016**



Gambar 2. Heatmap Korelasi Variabel Penelitian

Analisis korelasi dilakukan untuk mengetahui hubungan linear antarvariabel penelitian sebelum proses pemodelan menggunakan algoritma Random Forest Regression. Hasil analisis pada Tabel 3 dan Gambar 2 menunjukkan bahwa hubungan antara variabel independen dengan average_rating relatif rendah. Variabel ratings_count dan work_ratings_count memiliki koefisien korelasi sebesar 0,045, sedangkan work_text_reviews_count memiliki korelasi sebesar 0,007. Variabel books_count menunjukkan korelasi negatif sebesar −0,070, sementara original_publication_year memiliki korelasi positif yang sangat lemah sebesar 0,016 terhadap average_rating.

Di sisi lain, hubungan yang sangat kuat ditemukan antara ratings_count dan work_ratings_count dengan nilai koefisien korelasi sebesar 0,995, menunjukkan bahwa kedua variabel menggambarkan karakteristik popularitas buku yang hampir sama. Selain itu, work_text_reviews_count memiliki hubungan yang kuat dengan work_ratings_count (0,807) dan ratings_count (0,780), sedangkan books_count memiliki hubungan sedang dengan kedua variabel tersebut. Secara umum, hasil analisis menunjukkan bahwa sebagian besar hubungan linear terhadap average_rating relatif lemah sehingga diperlukan pemodelan menggunakan algoritma Random Forest Regression yang mampu mengidentifikasi hubungan nonlinier antarvariabel.

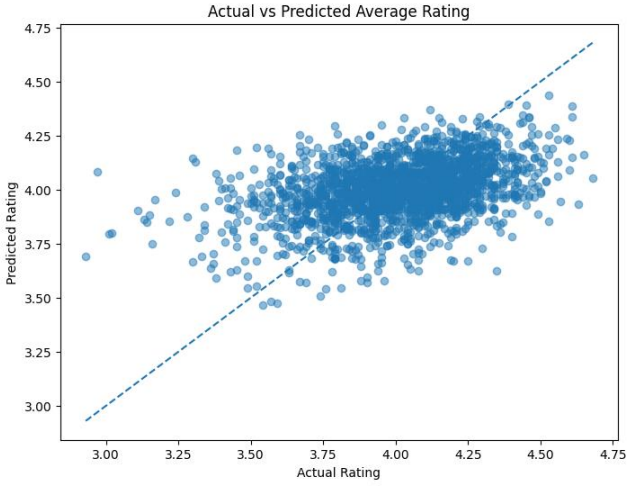
3. Evaluasi Model Random Forest Regression

Setelah proses preprocessing, data dibagi menjadi 80% data pelatihan (training set) dan 20% data pengujian (testing set). Selanjutnya dilakukan pemodelan menggunakan algoritma Random Forest Regression untuk menganalisis faktor-faktor yang memengaruhi nilai average_rating berdasarkan variabel metadata buku. Kinerja model dievaluasi menggunakan tiga metrik, yaitu Coefficient of Determination (R^2), Mean Absolute Error (MAE), dan Root Mean Square Error (RMSE). Hasil evaluasi model disajikan pada Tabel 4.

Tabel 4. Hasil Evaluasi Model Random Forest Regression

Metrik	Nilai
R^2	**0,1424**
MAE	**0,1791**
RMSE	**0,2292**

Berdasarkan Tabel 4 diperoleh nilai R^2 sebesar 0,1424, yang menunjukkan bahwa model mampu menjelaskan sekitar 14,24% variasi nilai average_rating berdasarkan variabel yang digunakan dalam penelitian. Nilai R^2 sebesar 0,1424 menunjukkan bahwa metadata bibliografis yang digunakan hanya mampu menjelaskan sebagian kecil variasi average_rating. Hal ini mengindikasikan bahwa terdapat faktor lain di luar metadata numerik, seperti genre, reputasi penulis, kualitas isi buku, dan sentimen ulasan yang kemungkinan memiliki pengaruh lebih besar. Nilai MAE sebesar 0,1791 menunjukkan bahwa rata-rata kesalahan absolut prediksi model sebesar 0,1791 poin dari nilai *rating* aktual, sedangkan nilai RMSE sebesar 0,2292 menunjukkan bahwa rata-rata kesalahan prediksi model masih berada pada kisaran yang relatif rendah terhadap skala penilaian *Goodreads*.



Gambar 3. Scatter Plot Actual vs Predicted

Selain menggunakan metrik evaluasi R^2, MAE, dan RMSE, kinerja model juga divisualisasikan melalui scatter plot antara nilai aktual dan nilai prediksi average_rating sebagaimana ditunjukkan pada Gambar 3. Garis diagonal pada grafik merepresentasikan kondisi ideal ketika nilai prediksi sama dengan nilai aktual.

Berdasarkan Gambar 3, sebagian besar titik data terkonsentrasi pada rentang nilai 3,75 hingga 4,25 dan cenderung mengikuti arah garis diagonal. Hal ini menunjukkan bahwa model mampu menangkap pola umum distribusi rating buku pada dataset Goodreads. Namun demikian, penyebaran titik yang masih cukup lebar serta kecenderungan model memprediksi nilai di sekitar

rata-rata menunjukkan bahwa kemampuan prediksi model masih terbatas. Fenomena ini mengindikasikan bahwa metadata bibliografis yang digunakan dalam penelitian hanya mampu menjelaskan sebagian variasi nilai average_rating. Temuan tersebut sejalan dengan nilai koefisien determinasi (R^2) sebesar 0,1424 yang menunjukkan bahwa masih terdapat faktor lain di luar variabel penelitian yang memengaruhi rating buku, seperti karakteristik isi buku, genre, reputasi penulis, maupun preferensi pembaca.

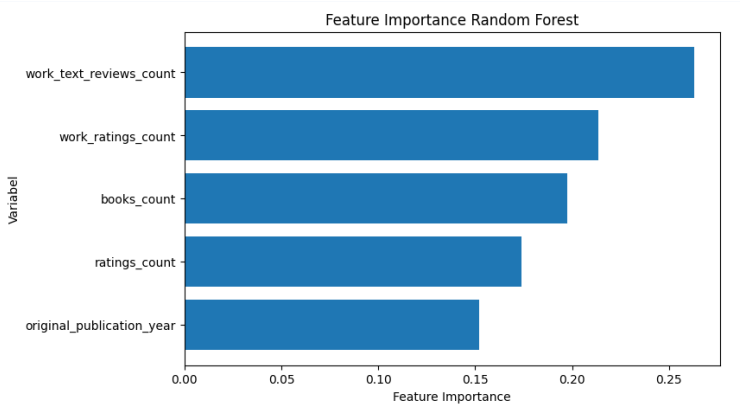
4. Analisis Feature Importance

Untuk mengetahui kontribusi masing-masing variabel terhadap hasil prediksi, dilakukan analisis feature importance pada model Random Forest Regression. Nilai *feature importance* menunjukkan besarnya pengaruh relatif setiap variabel dalam membentuk keputusan model.

Tabel 5. Feature Importance

Peringkat	Variabel	Feature Importance
1	Work Text Reviews Count	**0,2631**
2	Work Ratings Count	**0,2135**
3	Books Count	**0,1975**
4	Ratings Count	**0,1738**
5	Original Publication Year	**0,1521**

Selain disajikan dalam bentuk tabel, nilai *feature importance* divisualisasikan menggunakan grafik batang horizontal sebagaimana ditunjukkan pada Gambar 4.



Gambar 4. Feature Importance Variabel pada Model Random Forest Regression

Berdasarkan Tabel 5 dan Gambar 4 diketahui bahwa variabel work_text_reviews_count memiliki nilai *feature importance* tertinggi yaitu 0,2631, diikuti oleh work_ratings_count sebesar 0,2135, books_count sebesar 0,1975, ratings_count sebesar 0,1738, dan original_publication_year sebesar 0,1521. Hasil tersebut menunjukkan bahwa seluruh variabel memberikan kontribusi terhadap proses prediksi nilai average_rating, meskipun dengan tingkat

kepentingan yang berbeda. Secara umum, hasil pemodelan menunjukkan bahwa algoritma Random Forest Regression mampu mengidentifikasi tingkat kepentingan setiap variabel metadata dalam memprediksi nilai rata-rata *rating* buku. Hasil ini menjadi dasar untuk dilakukan interpretasi lebih lanjut mengenai faktor-faktor yang memengaruhi *rating* buku pada bagian pembahasan.

PEMBAHASAN

Hasil analisis data menggunakan teknik *data mining* melalui algoritma *Random Forest Regression* memberikan perspektif baru yang mendalam mengenai faktor-faktor metadata bibliografis yang memengaruhi nilai rata-rata rating (*average_rating*) buku pada platform *Goodreads*. Berdasarkan temuan awal melalui analisis korelasi Pearson, seluruh variabel independen secara linear menunjukkan hubungan yang sangat lemah terhadap *average_rating*. Hal ini mengonfirmasi argumen Baskara et al. (2025) bahwa dinamika data berskala besar sering kali tidak bersifat linear sederhana, sehingga memerlukan pendekatan lanjut seperti *ensemble machine learning* untuk mengidentifikasi pola hubungan yang lebih kompleks.

1. Analisis Kontribusi Variabel Berdasarkan *Feature Importance*

Melalui implementasi algoritma *Random Forest Regression*, kontribusi masing-masing variabel metadata dapat dipetakan secara non-linear melalui metrik *feature importance*. Hasil pemodelan menempatkan jumlah ulasan teks (*work_text_reviews_count*) sebagai faktor yang paling berpengaruh terhadap pembentukan nilai rata-rata rating dengan nilai 0,2631. Temuan ini memperluas analisis dari Kusumaningtyas & Ramdhani (2025) yang menyatakan bahwa ulasan teks di media sosial bukan sekadar elemen pelengkap bibliografis, melainkan representasi sentimen mendalam dari interaksi aktif pengguna. Tingginya nilai *feature importance* pada ulasan teks menunjukkan bahwa buku yang mampu memicu diskusi tekstual yang intens cenderung mengalami fluktuasi atau penguatan pada nilai rating rata-ratanya. Hal ini sejalan dengan pandangan Khadijah et al. (2025) yang melihat *Goodreads* sebagai ruang interaksi terbuka yang mengonstruksi resepsi pembaca secara langsung dan dinamis.

Faktor kedua dan keempat terbesar ditempati oleh *work_ratings_count* (0,2135) dan *ratings_count* (0,1738). Tingginya pengaruh metrik volume penilaian ini membuktikan bahwa popularitas buku berbanding lurus dengan intensitas penilaian kuantitatif yang diterimanya. Secara statistik, analisis korelasi juga menunjukkan hubungan linear yang sangat kuat (0,995) antara kedua variabel ini. Kedua variabel tetap dipertahankan karena Random Forest relatif tahan terhadap multikolinearitas dibanding regresi linear. Temuan ini mengindikasikan adanya efek akumulasi massa (*social proof*); buku yang memiliki jumlah pemberi rating tinggi akan terus

menarik pengguna lain untuk membaca dan memberikan penilaian, yang pada akhirnya memengaruhi stabilitas *average_rating* buku tersebut pada angka rata-rata 4,002.

Di sisi lain, jumlah edisi buku (*books_count*) menempati urutan ketiga dengan tingkat kepentingan sebesar 0,1975, serta memiliki korelasi negatif terkecil yaitu −0,070. Hubungan negatif yang lemah ini menarik untuk dicermati; semakin banyak edisi suatu buku yang diterbitkan (misalnya adaptasi, cetak ulang, atau terjemahan), terdapat kecenderungan kecil penurunan nilai rata-rata rating. Hal ini diduga terjadi karena variasi kualitas antaredisi atau distribusi penilaian yang semakin tersebar meluas ke segmen pembaca yang lebih heterogen.

Terakhir, tahun publikasi pertama (*original_publication_year*) memiliki tingkat kepentingan terendah yaitu 0,1521. Ini mengisyaratkan bahwa faktor kebaruan waktu rilis suatu buku tidak menjadi penentu utama apakah buku tersebut akan mendapatkan apresiasi tinggi atau rendah dari komunitas pembaca digital saat ini.

2. Perbandingan dengan Penelitian Sebelumnya dan Kontribusi Ilmu Pengetahuan

Penerapan *data mining* dalam bidang Perpustakaan dan Sains Informasi sejauh ini masih didominasi oleh orientasi praktis, seperti optimalisasi penempatan koleksi fisik menggunakan metode *association rule* oleh Listianto et al. (2024) , atau pengembangan algoritma penyaringan otomatis (*collaborative filtering* dan *content-based filtering*) untuk sistem rekomendasi digital (Chauhan, 2021 dalam Galuh et al., 2025; Indah et al., 2026). Berbeda dengan studi-studi tersebut, penelitian ini menawarkan kebaruan (*novelty*) analitis dengan membongkar karakteristik intrinsik metadata bibliografis di balik performa sebuah buku.

Model regresi yang dihasilkan memiliki nilai Coefficient of Determination (R^2) sebesar 0,1424. Hal ini berarti variasi metadata bibliografis kuantitatif dalam dataset mampu menjelaskan 14,24% perubahan nilai *average_rating*. Angka ini memberikan kontribusi teoretis penting bagi ilmu informasi: *apresiasi atau rating sebuah buku pada platform digital tidak hanya ditentukan oleh kualitas teks atau konten semata, tetapi juga dipengaruhi secara signifikan oleh struktur interaksi digital di sekitarnya (seperti jumlah ulasan teks dan akumulasi rating)*. Sisa variasi nilai sebesar 85,76% kemungkinan besar dipengaruhi oleh faktor-faktor ekstrinsik lain yang tidak terekam dalam metadata kuantitatif baku, seperti popularitas penulis, tren genre, atau kampanye pemasaran digital.

3. Implikasi Praktis bagi Layanan Informasi dan Perpustakaan

Dalam konteks *evidence-based librarianship* dan transformasi digital, temuan ini memberikan landasan empiris yang kuat bagi pengelola perpustakaan dan lembaga dokumentasi untuk menerapkan pengambilan keputusan berbasis data (*data-driven decision making*).

1. **Pengembangan Koleksi Berbasis Interaksi (*Engagement-Driven Procurement*)**

   Pustakawan disarankan tidak hanya melihat angka sirkulasi peminjaman tradisional, tetapi juga mengintegrasikan metrik *work_text_reviews_count* dan *work_ratings_count* dari platform global seperti *Goodreads* sebagai indikator dalam memproyeksikan pengadaan buku baru. Buku dengan volume ulasan teks yang tinggi terbukti memicu keterlibatan emosional pembaca, sehingga berpotensi tinggi untuk menjadi koleksi yang diminati di perpustakaan.

2. **Evaluasi dan Kurasi Layanan Rekomendasi**

   Perpustakaan digital dapat mengintegrasikan bobot *feature importance* dari penelitian ini ke dalam sistem kurasi mereka. Rekomendasi buku tidak lagi hanya mengandalkan popularitas buta, tetapi mengombinasikan faktor *engagement* teks (ulasan) untuk menyajikan daftar buku berkualitas tinggi yang terbukti memuaskan kebutuhan pemustaka.

Secara keseluruhan, pemanfaatan ekosistem *big data* melalui metodologi *data mining* ini berhasil mengisi celah riset (*research gap*) yang ada , sekaligus membuktikan bahwa pengelolaan informasi modern harus adaptif terhadap pergeseran perilaku pembaca di era digital.

KESIMPULAN

Penelitian ini berhasil menjawab tujuan utama dalam mengidentifikasi dan menganalisis faktor-faktor metadata bibliografis yang memengaruhi nilai rata-rata rating buku pada platform Goodreads menggunakan algoritma *Random Forest Regression*. Hasil analisis data dan diskusi menyimpulkan bahwa apresiasi pembaca dalam bentuk rating digital tidak terbentuk oleh hubungan linear sederhana, melainkan berdasarkan nilai feature importance pada model Random Forest Regression, work_text_reviews_count dan work_ratings_count merupakan variabel yang memberikan kontribusi terbesar dalam prediksi average_rating. Namun, rendahnya nilai R^2 menunjukkan bahwa masih terdapat banyak faktor lain di luar metadata numerik yang memengaruhi rating buku. Temuan baru ini memberikan kontribusi teoretis penting bagi perkembangan sains informasi dan analisis perilaku pengguna digital, di mana persepsi kualitas atau kepuasan terhadap sebuah buku di era *big data* terbukti berbanding lurus dengan intensitas diskusi tekstual dan keterlibatan komunitas yang terbangun di sekitar buku tersebut, bukan sekadar ditentukan oleh faktor kebaruan waktu publikasi.

Berdasarkan keterbatasan model yang dihasilkan dalam penelitian ini, disarankan bagi penelitian selanjutnya untuk memperluas cakupan variabel dengan mengintegrasikan teknik *text mining* atau analisis sentimen langsung terhadap konten dari ulasan teks guna menangkap aspek kualitatif yang belum terwakili oleh metadata numerik. Selain itu, lembaga perpustakaan dan

penyedia layanan informasi disarankan untuk mulai mengadopsi metrik keterlibatan komunitas (*community engagement*) dari platform sosial sebagai indikator empiris mutakhir dalam merancang sistem rekomendasi buku otomatis dan mengevaluasi kebijakan pengembangan koleksi digital berbasis bukti.

REFERENSI

Amin, R., & Utami, A. S. F. (2025). Prediksi Nilai Ujian Berdasarkan Kebiasaan Siswa Menggunakan Algoritma Random Forest Regressor. *INFORMATION SYSTEM FOR EDUCATORS AND PROFESSIONALS : Journal of Information System*, *10*(2), 149. https://doi.org/10.51211/isbi.v10i2.3722

Baskara, A. A., Piranti, N. M., & Romdendine, M. F. (2025). FRAMEWORK DATA MINING : SEBUAH SURVEI. *Jurnal Mahasiswa Teknik Informatika)*, *9*(3). https://doi.org/doi.org/10.36040/jati.v9i3.13803

Dotulong, G. W. I., Kawuwung, P. C. E., & Santa, K. (2026). PEMANFAATAN BIG DATA DALAM MENDUKUNG PENGAMBILAN KEPUTUSAN BERBASIS DATA DI BERBAGAI BIDANG: STUDI LITERATUR. *Jurnal Pendidikan Teknologi Informasi Dan Komunikasi*, *6*(1), 234. https://calamus.id/index.php/edutik/article/view/288

Faradillah, Alie, M. F., & Mariana, S. (2025). Workshop Pengolahan Data Berbasis Data Mining dengan RapidMiner bagi Mahasiswa Tingkat Akhir. *Jurnal Abdimas Mandiri*, *9*(2), 298–307. https://doi.org/10.36982/jam.v9i2.5542

Firdaus, D. (2017). Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer. *Jurnal Format*, *6*, 2089–5615. https://www.academia.edu/download/94065287/224659-penggunaan-data-mining-dalam-kegiatan-si-f3afe53d.pdf

Galuh, A. C. A., Firliana, R., & Ristyawan, A. (2025). Penerapan Algoritma Apriori untuk Mengidentifikasi Pola Peminjaman Buku Pada Perpustakaan Mas Trip Kabupaten Kediri. *JSITIK: Jurnal Sistem Informasi Dan Teknologi Informasi Komputer*, *4*(1), 49–60. https://doi.org/10.53624/jsitik.v4i1.723

Indah, A. N., Fitriyana, A., & Yeni, H. (2026). SISTEM INFORMASI MANAJEMEN PERPUSTAKAAN BERBASIS AI MENGGUNAKAN COLLABORATIVE FILTERING UNTUK REKOMENDASI BUKU OTOMATIS DI PERPUSTAKAAN MAMUJU. *JUKONI: Jurnal Ilmu Ekonomi Dan Bisnis*, (1). https://doi.org/doi.org/10.70134/jukoni.v3i1.1167

Khadijah, Quraisy, M., & Djaliel, M. A. (2025). Respon Pembaca pada Situs Goodreads terhadap Novel Kita Pergi Hari Ini Karya Ziggy Z. *DEIKTIS: Jurnal Pendidikan Bahasa Dan Sastra*, *5*(4), 2025. https://dmi-journals.org/deiktis/index

Kusumaningtyas, J. A., & Ramdhani, D. A. G. (2025). Pemanfaatan Algoritma Text Mining untuk Pengelolaan Sistem Koleksi Perpustakaan Melalui Ulasan Buku di Media Sosial. *Journal of Librarianship and Information Science*, *5*(1), 43–50. https://doi.org/doi.org/10.20414/light.v5i1.11381

Listianto, A. B., Irma, A., & Ali, I. (2024). PEMANFAATAN DATA MINING UNTUK PENEMPATAN BUKU DI PERPUSTAKAAN MENGGUNAKAN METODE ASSOCIATION RULE. *Jurnal Mahasiswa Teknik Informatika*, *8*(2). https://www.ejournal.itn.ac.id/index.php/jati/article/download/8447/5347

Nurina, L., Sudarmanto, E., Susanto, E., Utami, R., & Ananda, S. (2024). Nusantara Computer and Design Review Integrasi Big Data dan Kecerdasan Buatan: Potensi dan Tantangan Menurut Tinjauan Literatur Sistematis. *NCDR*, *2*(1), 1–6. https://doi.org/10.55732/ncdr.v2i1.1204

Suryantari, P. A., Muttaqin, F., & Rahajoe, A. D. (2026). PERAN BIG DATA DALAM MANAJEMEN DATA DAN INFORMASI SEBAGAI SISTEM PENDUKUNG KEPUTUSAN (SYSTEMATIC LITERATURE REVIEW). *Jurnal Informatika Dan Teknik Elektro Terapan*, *14*(1). https://doi.org/10.23960/jitet.v14i1.8899

Tupan. (2024). Perkembangan Penelitian Penggunaan Artificial Intelligence di Perpustakaan Berbasis Data Scopus. *Media Pustakawan*, *31*(3), 277–290. https://doi.org/10.37014/medpus.v31i3.5316

Worrall, A. (2019). “Connections above and beyond”: Information, translation, and community boundaries in LibraryThing and Goodreads. *Journal of the Association for Information Science and Technology*, *70*(7), 742–753. https://doi.org/10.1002/asi.24153