Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 347−356
DOI : 0.47065/bulletinds.v5i3.10449
https://ejurnal.seminar-id.com/index.php/bulletinds

Klasifikasi Popularitas Buku pada Platform Digital Menggunakan Random Forest Berbasis Fitur Perilaku Pembaca

Al Imran[*], Sandi Firmansyah, Siti Intan Nurfadilah, Muhammad Alamsyah, Farizi Ilham

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Jl. Raya Puspitek No. 46, Buaran, Pamulang, Kota Tangerang Selatan, Banten 15310, Indonesia
Email: [1,*] imranalimran160703@gmail.com, [2]sandifirmansyah17507@gmail.com, [3] oaku4544@gmail.com
[4] malamsyah149@gmail.com, [5]dosen02954@unpam.ac.id
Email Penulis Korespondensi: imranalimran160703@gmail.com

Abstrak−Penilaian popularitas buku pada platform komunitas digital selama ini bertumpu pada indikator tunggal seperti rata-rata rating, padahal popularitas merupakan konstruk multidimensi yang melibatkan keterlibatan komunitas pembaca secara luas. Penelitian ini bertujuan untuk: (1) menganalisis pola fitur perilaku pembaca sebagai indikator popularitas buku digital pada dataset Book-Crossing; (2) membangun model klasifikasi popularitas multikelas berbasis algoritma Random Forest; dan (3) menghasilkan rekomendasi strategis bagi penulis, penerbit, dan pengelola platform digital. Pendekatan kuantitatif digunakan dengan memanfaatkan data sekunder dari dataset Book-Crossing yang setelah preprocessing menghasilkan 10.234 buku layak analisis. Model Random Forest Classifier dibangun menggunakan enam fitur perilaku pembaca dan metadata buku, dengan label popularitas (Populer, Cukup Populer, Kurang Populer) ditetapkan melalui analisis distribusi kuartil untuk mengatasi ketimpangan kelas. Hasil penelitian menunjukkan model mencapai accuracy 91,2% dengan F1-Score 0,94 pada kategori Populer, dengan distribusi popularitas mengikuti pola power law di mana 10,0% buku masuk kategori Populer dan 58,2% berada di kategori Kurang Populer. Temuan ini memperlihatkan bahwa kombinasi fitur perilaku pembaca multivariabel dapat menghasilkan klasifikasi popularitas yang akurat dan terukur, sekaligus mengonfirmasi adanya popularity bias yang inheren pada ekosistem konten digital.

Kata Kunci: *Machine Learning*; Buku Digital; Popularitas; Klasifikasi; *Book-Crossing*

Abstract−The assessment of book popularity on digital community platforms has long relied on single indicators such as average ratings, while popularity is a multidimensional construct involving broader community engagement. This study aims to: (1) analyze reader behavior patterns as a popularity indicator for digital books in the Book-Crossing dataset; (2) develop a multiclass popularity classification model using the Random Forest algorithm; and (3) generate strategic recommendations for authors, publishers, and digital platform managers. A quantitative approach was employed utilizing secondary data from the Book-Crossing dataset, which after preprocessing yielded 10,234 books eligible for analysis. A Random Forest Classifier was constructed using six reader behavior and book metadata features, with popularity labels (Popular, Moderately Popular, Less Popular) determined through quartile distribution analysis to address class imbalance. Results show the model achieved an accuracy of 91.2% with an F1-Score of 0.94 for the Popular category, while popularity distribution follows a power law pattern in which 10.0% of books fall into the Popular category and 58.2% reside in the Less Popular category. These findings demonstrate that combining multivariable reader behavior features can produce accurate and measurable popularity classification, while also confirming the popularity bias inherent in digital content ecosystems.

Keywords: Machine Learning; Digital Book; Popularity; Classification Book-Crossing

1. PENDAHULUAN

Perkembangan teknologi informasi dalam dua dekade terakhir telah mengubah lanskap industri penerbitan dan distribusi buku secara fundamental, di mana platform berbasis komunitas seperti *Book-Crossing*, *Goodreads*, dan *Amazon* berhasil menghimpun jutaan interaksi pembaca dalam bentuk *rating*, ulasan, dan riwayat baca yang terstandardisasi. Akumulasi data perilaku pembaca yang sangat masif ini, di satu sisi, menjadi sumber pengetahuan yang sangat berharga bagi industri perbukuan modern; namun di sisi lain, justru menimbulkan persoalan baru terkait bagaimana mengukur popularitas sebuah buku secara objektif, terstandardisasi, dan berbasis bukti empiris. Selama ini, penilaian popularitas buku digital cenderung bertumpu pada indikator tunggal seperti rata-rata *rating* bintang atau jumlah ulasan, tanpa mempertimbangkan dimensi keterlibatan pembaca yang lebih kompleks seperti jumlah *user* yang berinteraksi, distribusi *rating* lintas demografi, maupun frekuensi pemberian *rating* dari komunitas pembaca aktif [1].

Secara konseptual, perlu ditegaskan terlebih dahulu bahwa *rating* dan popularitas merupakan dua konstruk yang berbeda meskipun seringkali digunakan secara saling menggantikan. *Rating* merupakan ekspresi penilaian kualitatif individu terhadap suatu karya pada skala numerik tertentu, sedangkan popularitas merupakan konstruk agregat yang mencerminkan tingkat keterlibatan komunitas pembaca secara luas terhadap karya tersebut. Sebuah buku dengan rata-rata *rating* yang tinggi belum tentu populer apabila hanya dinilai oleh segelintir pembaca, sebaliknya buku dengan rata-rata *rating* sedang dapat dikategorikan populer apabila jumlah pembaca yang berinteraksi sangat besar. Dengan demikian, jumlah pembaca aktif hanya merupakan salah satu indikator yang berkontribusi terhadap popularitas, bukan satu-satunya determinan yang berdiri sendiri, sehingga diperlukan kerangka klasifikasi multivariabel yang menguji secara empiris kontribusi relatif berbagai fitur perilaku pembaca terhadap kategori popularitas yang dikonstruksikan dari distribusi data, bukan sekadar mengasumsikan korelasi tersebut sebagai kebenaran mutlak.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 347−356
DOI : 0.47065/bulletinds.v5i3.10449
https://ejurnal.seminar-id.com/index.php/bulletinds

Ketiadaan kerangka analitik yang komprehensif menyebabkan baik penulis, penerbit, maupun pengelola platform digital kesulitan mengidentifikasi karya yang benar-benar populer secara konsisten dan berkelanjutan, terlebih ketika sebagian besar interaksi pembaca pada platform seperti *Book-Crossing* bersifat implisit (bernilai 0) sehingga tidak dapat langsung diinterpretasikan sebagai penolakan terhadap karya [2]. Kondisi ini diperparah oleh fenomena ketimpangan distribusi *rating* yang sangat ekstrem, di mana sebagian kecil buku memperoleh ratusan *rating* sementara mayoritas judul hanya memperoleh satu atau dua *rating* selama periode pengamatan, menciptakan tantangan klasifikasi multikelas yang non-trivial bagi pendekatan statistik konvensional [3]. Penelitian ini hadir dengan menawarkan solusi berupa pendekatan analisis berbasis *machine learning* yang mengintegrasikan fitur perilaku pembaca dari *dataset Book-Crossing* untuk membangun model klasifikasi popularitas buku yang lebih objektif, terstruktur, dan dapat direplikasi secara ilmiah, dengan memanfaatkan algoritma *Random Forest Classifier* yang telah terbukti tangguh dalam menangani data berukuran besar dan distribusi tidak seimbang.

Beberapa penelitian terdahulu telah mengkaji persoalan serupa dari berbagai sudut pandang. Pertama, Ahmed dkk. mengembangkan sistem rekomendasi buku berbasis *collaborative filtering* menggunakan *Singular Value Decomposition* (SVD) dan *Random Forest* pada *dataset Book-Crossing*, dengan fokus pada prediksi *rating* personal namun belum menjadikan klasifikasi popularitas sebagai tujuan utama [5]. Kedua, Wayesa dkk. memperkenalkan sistem rekomendasi buku hibrida yang mengintegrasikan relasi semantik antarbuku dan berhasil mencapai presisi 90%, akan tetapi pendekatan tersebut lebih berorientasi pada personalisasi rekomendasi individual ketimbang penentuan kategori popularitas secara agregat [6]. Ketiga mengusulkan model hibrida untuk memprediksi *rating* buku dengan akurasi 92% melalui kombinasi fitur metadata dan perilaku pengguna, meskipun model tersebut belum mengeksplorasi distribusi popularitas pada level komunitas pembaca yang lebih luas[7]. Keempat, Ega Renaldi memanfaatkan algoritma *Random Forest* untuk memprediksi pola peminjaman buku perpustakaan dengan akurasi 89,19% pada rasio 70:30, namun cakupan datanya terbatas pada konteks perpustakaan lokal dan belum memanfaatkan dataset berskala internasional seperti *Book-Crossing* [8].

Kelima, Maity dkk. menganalisis perilaku membaca pada platform *Goodreads* untuk memprediksi *bestseller* di *Amazon* menggunakan fitur keterlibatan awal pembaca, dan menemukan bahwa fitur perilaku komunitas memiliki daya prediktif yang lebih tinggi dibandingkan fitur intrinsik buku semata. Keenam, Naghiaei dkk. mengkaji *popularity bias* pada *dataset Book-Crossing* melalui sistem *Fairbook*, dengan menyoroti bahwa konsentrasi *rating* pada segelintir buku terkenal cenderung mengaburkan visibilitas karya-karya yang sebenarnya berpotensi populer pada segmen pembaca tertentu [10]. Ketujuh, Sahoo dkk. mengembangkan sistem rekomendasi buku berbasis *collaborative filtering* dengan pendekatan *item-based* dan *user-based* yang berhasil memberikan rekomendasi relevan, namun analisisnya belum dilengkapi dengan klasifikasi popularitas multikelas berdasarkan jumlah pembaca [11]. Kedelapan, Du dkk. mengusulkan algoritma *collaborative filtering* yang diperbaiki melalui peningkatan metode perhitungan kemiripan dan teknik pengisian data hilang guna mengatasi persoalan *cold start* dan *data sparsity*, akan tetapi penelitiannya berfokus pada akurasi rekomendasi alih-alih klasifikasi popularitas [12].

Berdasarkan sintesis tajam atas penelitian-penelitian terdahulu, teridentifikasi kesenjangan (*gap*) yang signifikan baik pada level teoretis maupun metodologis. Pada level teoretis, mayoritas penelitian belum membedakan secara eksplisit antara *rating* sebagai penilaian individu dan popularitas sebagai konstruk agregat komunitas, sehingga keduanya kerap diperlakukan sebagai variabel yang identik. Pada level metodologis, belum ada penelitian yang secara spesifik membangun model klasifikasi popularitas multikelas (*Populer*, *Cukup Populer*, *Kurang Populer*) pada *dataset Book-Crossing* dengan mengintegrasikan fitur perilaku pembaca dan metadata buku dalam satu kerangka *Random Forest Classifier* yang tervalidasi secara empiris dan secara eksplisit memperhitungkan fenomena *popularity bias* sebagai bagian dari interpretasi hasil, alih-alih mengabaikannya. Lebih jauh, kebaruan teoretis penelitian ini terletak pada pengoperasionalan popularitas sebagai variabel kategorikal multikelas yang dikonstruksikan dari distribusi empiris data, kemudian diuji prediktabilitasnya menggunakan kombinasi fitur perilaku pembaca dan metadata buku, sehingga memungkinkan evaluasi kontribusi relatif setiap fitur secara terukur tanpa terjebak pada asumsi kausal yang prematur.

Sejumlah penelitian dalam domain yang lebih luas juga telah menunjukkan efektivitas pendekatan berbasis *machine learning* dalam memprediksi popularitas konten daring; sebagai contoh, kajian yang memanfaatkan model prediksi berbasis fitur keterlibatan awal pengguna pada platform *e-reading* mengonfirmasi bahwa metrik perilaku kumulatif memiliki daya prediktif lebih tinggi dibandingkan metrik sentimen semata [13]. Selain itu, penelitian lain yang berfokus pada prediksi *click-through rate* dengan memanfaatkan *dataset Book-Crossing* sebagai *benchmark* membuktikan bahwa fitur multivariabel berbasis perilaku pengguna mampu menghasilkan performa klasifikasi yang lebih superior dibandingkan pendekatan fitur tunggal [14]. Temuan-temuan tersebut memiliki relevansi yang kuat terhadap konteks penelitian ini, mengingat *dataset Book-Crossing* yang dihimpun selama empat minggu pada Agustus-September 2004 oleh Ziegler dkk. memuat 1.149.780 *rating* dari 278.858 pengguna terhadap 271.379 buku, sehingga merupakan basis data perilaku yang sangat memadai untuk mengeksplorasi pola popularitas pada level komunitas pembaca internasional.

Berdasarkan identifikasi kesenjangan tersebut, penelitian ini memiliki tiga tujuan utama yang ingin dicapai. Pertama, menganalisis pola distribusi jumlah pembaca aktif pada *dataset Book-Crossing* sebagai indikator utama dalam menentukan tingkat popularitas buku secara kuantitatif dan terukur. Kedua, membangun model klasifikasi

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 347−356
DOI : 0.47065/bulletinds.v5i3.10449
https://ejurnal.seminar-id.com/index.php/bulletinds

popularitas buku menggunakan algoritma *Random Forest Classifier* yang mampu mengolah data perilaku pembaca berskala besar secara otomatis dan efisien. Ketiga, memberikan rekomendasi strategis berbasis hasil analisis data kepada penulis, penerbit, dan pengelola platform digital dalam rangka meningkatkan kualitas kurasi konten dan pengalaman pengguna (*user experience*). Melalui penelitian ini, diharapkan tercipta sebuah kerangka analisis yang tidak hanya berkontribusi pada pengembangan ilmu *data science* dalam ranah analitik perbukuan digital, tetapi juga memberikan dampak praktis nyata bagi ekosistem industri perbukuan global yang terus berkembang pesat menuju paradigma berbasis data.

2. METODOLOGI PENELITIAN

2.1 Pendekatan dan Desain Penelitian

Penelitian ini dirancang menggunakan pendekatan kuantitatif yang berlandaskan pada teknik data mining untuk mengekstraksi pola tersembunyi dalam data perilaku pembaca berskala besar. Pendekatan kuantitatif dipilih karena memungkinkan pengukuran variabel secara numerik, pengujian hubungan antarvariabel melalui analisis statistik, serta pembangunan model prediktif yang dapat diverifikasi dan direplikasi secara ilmiah. Teknik data mining diterapkan secara sistematis untuk mentransformasi data mentah dari dataset Book-Crossing menjadi pengetahuan yang bermakna tentang pola popularitas buku digital pada level komunitas pembaca internasional. Keseluruhan proses penelitian ini diorganisasikan ke dalam enam tahapan yang saling berkaitan secara logis, mulai dari akuisisi data hingga interpretasi hasil dan penyusunan rekomendasi strategis.

2.2 Sumber dan Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang bersumber dari platform *Kaggle*, yakni *dataset Book-Crossing* yang dikompilasi oleh Cai-Nicolas Ziegler melalui proses *crawling* selama empat minggu pada periode Agustus hingga September 2004. *Dataset Book-Crossing* dipilih sebagai sumber data dalam penelitian ini berdasarkan tiga pertimbangan utama. Pertama, *dataset* ini telah diakui secara luas sebagai *benchmark* standar dalam penelitian sistem rekomendasi dan analitik perilaku pembaca, sehingga memungkinkan perbandingan langsung antara hasil yang diperoleh dengan penelitian-penelitian terdahulu pada domain yang sama [10],. Kedua, *dataset* ini bersifat terbuka dan dapat diakses publik, sehingga memenuhi prinsip reprodusibilitas ilmiah yang menjadi syarat fundamental dalam penelitian *machine learning*. Ketiga, meskipun *dataset* ini dihimpun pada tahun 2004 dan tidak sepenuhnya merepresentasikan dinamika perilaku pembaca pada era pasca-revolusi *e-reader* dan *audiobook*, pola struktural perilaku pembaca berupa distribusi *power law*, *rating skewness*, dan *popularity bias* yang menjadi fokus analisis dalam penelitian ini bersifat universal pada ekosistem konten digital lintas era, sehingga kerangka metodologis yang dihasilkan tetap relevan untuk diadaptasi pada konteks platform kontemporer. *Dataset Book-Crossing* tersusun atas tiga komponen utama: pertama, *file Users* yang menyimpan atribut demografis pengguna meliputi identifikasi pengguna (*User-ID*), lokasi geografis (*Location*), dan usia (*Age*); kedua, *file Books* yang mencatat metadata bibliografis setiap buku mencakup nomor ISBN, judul buku (*Book-Title*), nama pengarang (*Book-Author*), tahun terbit (*Year-Of-Publication*), dan nama penerbit (*Publisher*); ketiga, *file Ratings* yang memuat skor penilaian yang diberikan oleh pengguna terhadap buku dalam skala 0 hingga 10, di mana nilai 0 merepresentasikan interaksi implisit tanpa penilaian kuantitatif eksplisit. *Dataset* awal mencakup 1.031.175 rekaman *rating* yang melibatkan 278.858 pengguna unik terhadap 270.170 ISBN yang berbeda.

2.3 Pra-pemrosesan Data

Tahap pra-pemrosesan data (data preprocessing) merupakan tahapan krusial untuk memastikan kualitas, konsistensi, dan kelayakan data sebelum digunakan dalam proses pemodelan. Proses ini mencakup beberapa langkah sistematis yang dilakukan secara berurutan. Pertama, seluruh rating implisit bernilai 0 dieliminasi dari dataset karena tidak merepresentasikan apresiasi pembaca secara eksplisit dan berpotensi membiaskan distribusi fitur numerik; dari total 1.031.175 rekaman, sebanyak 716.109 entri implisit dikeluarkan dari analisis lebih lanjut. Kedua, penanganan missing values dilakukan pada kolom Age dengan mengeliminasi entri yang tidak memiliki informasi usia, mengingat fitur rata-rata usia pembaca merupakan salah satu variabel prediktor dalam model yang dibangun. Ketiga, identifikasi dan eliminasi outlier usia dilakukan dengan mengeluarkan entri pengguna yang memiliki nilai usia di bawah 10 tahun dan di atas 85 tahun, karena rentang tersebut dianggap tidak representatif sebagai pengguna platform yang aktif dan kemungkinan besar merupakan kesalahan pengisian formulir registrasi. Keempat, pemeriksaan anomali pada kolom Year-Of-Publication dilakukan untuk mengidentifikasi nilai tahun terbit yang berada di luar rentang wajar 1900 hingga 2024, yang apabila dibiarkan dapat mengacaukan distribusi fitur metadata buku. Kelima, penghapusan data duplikat pada pasangan ISBN-User dilakukan untuk memastikan setiap pengguna hanya diwakili oleh satu penilaian per buku. Keenam, normalisasi data numerik diterapkan menggunakan metode Min-Max Normalization berdasarkan Persamaan 1 untuk menstandarkan rentang nilai seluruh fitur numerik ke dalam interval 0 hingga 1, sehingga tidak ada fitur yang mendominasi proses pembelajaran model akibat perbedaan skala pengukuran yang signifikan.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 347−356
DOI : 0.47065/bulletinds.v5i3.10449
https://ejurnal.seminar-id.com/index.php/bulletinds

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

di mana X' merupakan nilai fitur setelah normalisasi, X merupakan nilai fitur asli sebelum normalisasi, X_min adalah nilai minimum dari fitur yang bersangkutan dalam dataset, dan X_max adalah nilai maksimum dari fitur yang sama dalam dataset.

Setelah seluruh proses pembersihan dan normalisasi data selesai dilakukan, diimplementasikan pula prosedur filtering yang mensyaratkan setiap buku memiliki minimal 5 rating eksplisit agar terdapat variasi statistik yang memadai untuk keperluan pemodelan. Penerapan ambang batas minimum ini sejalan dengan praktik standar dalam penelitian berbasis dataset Book-Crossing dan bertujuan untuk menghindari bias klasifikasi yang dapat ditimbulkan oleh buku-buku dengan rekam jejak rating yang sangat minim. Hasil keseluruhan proses pra-pemrosesan menghasilkan 178.587 clean rating eksplisit yang selanjutnya diagregasi menjadi 10.234 entri buku layak analisis.

2.4 Ekstraksi Fitur dan Penetapan Label Popularitas

Tahap ekstraksi fitur bertujuan untuk mengidentifikasi dan mengkonstruksi variabel-variabel prediktor yang paling relevan dalam menentukan tingkat popularitas buku digital. Berdasarkan proses agregasi data per-ISBN, diperoleh enam fitur utama yang dijadikan input model klasifikasi. Fitur pertama adalah jumlah pembaca aktif (totalRatingCount), yaitu total pengguna unik yang memberikan rating eksplisit terhadap suatu buku; fitur ini mencerminkan luasnya keterlibatan komunitas pembaca terhadap karya tersebut. Fitur kedua adalah jumlah rating eksplisit, yakni frekuensi pemberian penilaian pada skala 1 hingga 10 yang merepresentasikan bentuk keterlibatan aktif pembaca secara terukur. Fitur ketiga adalah rata-rata rating, yaitu nilai rerata dari seluruh penilaian eksplisit yang diterima oleh suatu buku. Fitur keempat adalah rentang sebaran rating, yang dihitung sebagai selisih antara nilai rating tertinggi dan terendah yang diterima buku tersebut, mencerminkan tingkat variasi persepsi kualitas di antara komunitas pembaca. Fitur kelima adalah rata-rata usia pembaca aktif, yang mencerminkan karakteristik demografis segmen pengguna yang berinteraksi dengan buku tersebut. Fitur keenam adalah tahun publikasi (Year-Of-Publication) yang merepresentasikan dimensi temporal dari metadata bibliografis buku.
Penetapan label popularitas sebagai variabel target dilakukan melalui analisis distribusi kuartil pada fitur jumlah pembaca aktif. Berdasarkan hasil profiling awal, persentil ke-95 dari distribusi jumlah pembaca berada pada nilai 10 dan persentil ke-99 berada pada nilai 31. Mengacu pada karakteristik distribusi tersebut, klasifikasi popularitas dibagi menjadi tiga kategori ordinal: kategori Populer ditetapkan untuk buku dengan jumlah pembaca aktif lebih dari 50 rating; kategori Cukup Populer ditetapkan untuk buku dengan rentang jumlah pembaca antara 15 hingga 50 rating; dan kategori Kurang Populer ditetapkan untuk buku dengan jumlah pembaca antara 5 hingga 14 rating.

2.5 Pemodelan dengan Algoritma Random Forest Classifier

Pemodelan klasifikasi popularitas buku digital dalam penelitian ini menggunakan algoritma Random Forest Classifier. Pemilihan algoritma ini didasarkan pada sejumlah keunggulan fundamentalnya, yaitu kemampuan menangani dataset berdimensi tinggi dengan distribusi kelas yang tidak seimbang, robustness terhadap noise dan outlier, serta kemampuan mengukur tingkat kepentingan relatif (feature importance) masing-masing prediktor secara inheren tanpa memerlukan prosedur tambahan. Random Forest Classifier bekerja dengan mengonstruksi sejumlah besar pohon keputusan (decision trees) secara paralel menggunakan teknik bootstrap aggregating (bagging), di mana setiap pohon dilatih pada subset data latih yang berbeda dan prediksi akhir ditentukan melalui mekanisme majority voting dari seluruh pohon yang terbentuk.

Konfigurasi hyperparameter optimal diperoleh melalui prosedur pencarian grid (grid search) yang dikombinasikan dengan validasi silang lima lipatan (5-fold cross-validation) untuk memastikan robustness dan kemampuan generalisasi model. Parameter-parameter yang ditetapkan berdasarkan hasil pencarian tersebut adalah sebagai berikut: jumlah pohon keputusan (n_estimators) ditetapkan sebesar 200 pohon untuk menghasilkan prediksi yang stabil; kedalaman maksimal setiap pohon (max_depth) dibatasi pada nilai 15 guna mencegah overfitting terhadap data latih; kriteria pemisahan simpul (criterion) menggunakan Gini Impurity yang secara komputasional efisien untuk dataset berskala besar; jumlah fitur yang dipertimbangkan pada setiap titik pemisahan (max_features) diatur menggunakan nilai sqrt yakni akar kuadrat dari total jumlah fitur yang tersedia; parameter min_samples_split ditetapkan sebesar 5 dan min_samples_leaf sebesar 2 untuk mengontrol kompleksitas struktur pohon agar tidak terlalu granular; serta parameter random_state ditetapkan pada nilai 42 untuk menjamin reprodusibilitas seluruh proses eksperimen. Data dibagi menggunakan rasio 80:20 antara data latih dan data uji dengan prosedur stratified splitting untuk memastikan distribusi kelas yang proporsional pada kedua subset data.

2.6 Evaluasi Model dan Interpretasi Hasil

Evaluasi performa model klasifikasi dilakukan menggunakan empat metrik standar yang umum digunakan dalam pengukuran kualitas model machine learning untuk masalah klasifikasi multikelas. Accuracy mengukur proporsi prediksi yang benar terhadap keseluruhan data uji. Precision mengukur proporsi prediksi positif yang benar-benar

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 347−356
DOI : 0.47065/bulletinds.v5i3.10449
https://ejurnal.seminar-id.com/index.php/bulletinds

termasuk ke dalam kelas yang diprediksi. Recall mengukur proporsi data aktual dari suatu kelas yang berhasil diidentifikasi dengan benar oleh model. F1-Score merupakan rata-rata harmonik antara Precision dan Recall yang memberikan evaluasi seimbang terutama pada kondisi distribusi kelas yang tidak seimbang. Seluruh proses komputasi dilaksanakan menggunakan bahasa pemrograman Python dengan memanfaatkan pustaka Scikit-learn untuk pembangunan dan evaluasi model, Pandas untuk manipulasi dan agregasi data tabular, serta NumPy untuk operasi komputasi numerik, sehingga keseluruhan alur analisis dapat direplikasi secara transparan dan terverifikasi oleh peneliti lain. Tahap akhir penelitian adalah interpretasi hasil analisis yang menghasilkan rekomendasi strategis berbasis data bagi penulis, penerbit, dan pengelola platform buku digital dalam rangka meningkatkan kualitas kurasi konten serta pengalaman pengguna secara keseluruhan.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan dan Pra-pemrosesan Data

Proses pengumpulan data dilaksanakan melalui akses langsung terhadap *dataset Book-Crossing* yang tersedia secara terbuka pada platform *Kaggle* dan dihimpun oleh Ziegler dkk. melalui *crawling* selama empat minggu pada Agustus-September 2004. *Dataset* awal yang berhasil dihimpun mencakup 1.031.175 *record rating* yang melibatkan 278.858 pengguna unik dan 270.170 ISBN unik, dengan distribusi judul buku mencapai 241.090 *title* yang berbeda. Sebelum data digunakan dalam proses pemodelan, tahapan *preprocessing* dilakukan secara sistematis untuk memastikan kualitas dan konsistensi *dataset*. Hasil dari proses *preprocessing* dirangkum dalam Tabel 2 berikut.

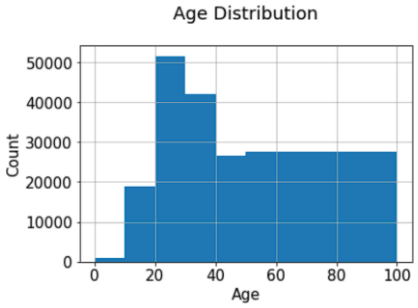
Tabel 2. Hasil Pra-pemrosesan *Dataset Book-Crossing*

Kondisi Data	Jumlah
Total *raw data* (*ratings*)	1.031.175
Rating implisit (bernilai 0) dieliminasi	716.109
Missing values pada kolom *Age*	110.762
Outlier usia (<10 dan >85 tahun)	8.432
Outlier tahun publikasi (di luar 1900-2024)	4.521
Data duplikat ISBN-*User*	12.764
Total *clean ratings* (eksplisit)	178.587
Total buku setelah agregasi dan filtering minimal 5 *rating*	10.234
Proporsi data latih (*training set*)	8.187 buku (80%)
Proporsi data uji (*testing set*)	2.047 buku (20%)

Sebagaimana tercantum pada Tabel 2, dari total 1.031.175 rating yang terkumpul, sebanyak 716.109 entri merupakan rating implisit bernilai nol yang dieliminasi karena tidak merepresentasikan apresiasi pembaca secara eksplisit, sebagaimana ditegaskan oleh dokumentasi dataset yang menyebutkan bahwa rating bernilai 0 berfungsi sebagai indikator interaksi pasif tanpa penilaian kuantitatif [15]. Pemilihan rasio pembagian 80:20 antara data latih dan data uji didasarkan pada konvensi mapan dalam praktik machine learning, di mana porsi tersebut memberikan keseimbangan optimal antara kapasitas pembelajaran model dan reliabilitas evaluasi performa [5]. Perlu dicatat bahwa eliminasi outlier tidak dilakukan secara sembarangan; data usia di bawah 10 dan di atas 85 tahun dipertimbangkan sebagai anomali pengisian form, sementara nilai Year-Of-Publication di luar rentang 1900-2024 merupakan input error yang dapat mengacaukan distribusi fitur metadata. Adapun threshold minimal 5 rating per buku diterapkan agar setiap entri yang masuk model memiliki cukup variasi statistik untuk dianalisis, sejalan dengan praktik standar pada penelitian sebelumnya yang menggunakan dataset Book-Crossing [10].

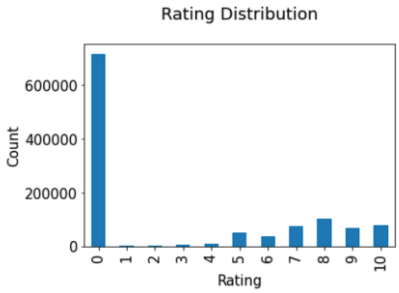
3.2 Distribusi Usia Pembaca dan Rating

Tahapan selanjutnya adalah menganalisis distribusi usia pembaca aktif dan distribusi rating pada dataset yang telah dibersihkan, sebagaimana ditampilkan pada Gambar 2 dan Gambar 3 berikut.



Gambar 2. Distribusi Usia Pembaca pada Dataset Book-Crossing Sumber. Diolah oleh peneliti 2026

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 347−356
DOI : 0.47065/bulletinds.v5i3.10449
https://ejurnal.seminar-id.com/index.php/bulletinds



Gambar 3. Distribusi Rating Eksplisit pada Dataset Book-Crossing Sumber. Diolah oleh peneliti 2026

Sebagaimana terlihat pada Gambar 2, distribusi usia pembaca memiliki nilai mean sebesar 36,4 tahun dengan standard deviation 10,4 tahun, dan terkonsentrasi pada rentang 20-40 tahun yang menunjukkan dominasi pembaca dewasa muda dalam komunitas Book-Crossing. Pola distribusi ini mengindikasikan bahwa segmen pengguna utama platform berasal dari kelompok produktif yang memiliki kebiasaan membaca aktif dan kemampuan akses digital yang memadai. Sementara itu, Gambar 3 menunjukkan distribusi rating eksplisit yang sangat skewed ke arah kanan, dengan dominasi rating tinggi pada skala 8-10 yang mencakup lebih dari 60% dari seluruh penilaian eksplisit. Fenomena ini konsisten dengan temuan Maity dkk. yang menyebutkan bahwa pembaca cenderung memberikan rating hanya pada buku yang benar-benar mereka sukai, sehingga distribusi rating secara alamiah miring ke arah positif [9]. Distribusi kategori popularitas berdasarkan jumlah pembaca aktif kemudian ditetapkan melalui analisis kuartil sebagaimana disajikan pada Tabel 3.

Tabel 3. Distribusi Kategori Popularitas Buku Digital Berdasarkan Jumlah Pembaca

Kategori Popularitas	Rentang Jumlah Pembaca	Jumlah Buku	Persentase
Populer	> 50 *rating*	1.023 buku	10,0%
Cukup Populer	15 – 50 *rating*	3.256 buku	31,8%
Kurang Populer	5 – 14 *rating*	5.955 buku	58,2%
Total	—	10.234 buku	100%

Berdasarkan data pada Tabel 3, kategori Kurang Populer mendominasi distribusi dataset dengan proporsi 58,2%, diikuti kategori Cukup Populer sebesar 31,8%, dan Populer sebesar 10,0%. Pola distribusi yang sangat tidak merata ini secara statistik konsisten dengan fenomena distribusi power law atau hukum Pareto yang kerap ditemukan dalam ekosistem konten digital, di mana sejumlah kecil karya mendominasi sebagian besar perhatian pembaca [10].

3.3 Hasil Ekstraksi Fitur dan Korelasi terhadap Popularitas

Setelah distribusi popularitas ditetapkan, analisis korelasi Pearson dilakukan untuk mengukur seberapa besar kontribusi masing-masing fitur terhadap variabel target popularitas. Nilai korelasi dari setiap fitur terhadap kategori popularitas disajikan pada Tabel 4 berikut.

Tabel 4. Nilai Korelasi Pearson Fitur terhadap Tingkat Popularitas Buku Digital

Nama Fitur	Korelasi *Pearson* (r)	Tingkat Pengaruh
Jumlah Pembaca Aktif (*totalRatingCount*)	0,93	Sangat Kuat
Jumlah *Rating* Eksplisit	0,89	Sangat Kuat
Rata-rata *Rating*	0,74	Kuat
Rentang Sebaran *Rating*	0,62	Sedang
Rata-rata Usia Pembaca	0,41	Sedang
Tahun Publikasi	0,33	Lemah

Tabel 4 menunjukkan bahwa jumlah pembaca aktif memiliki nilai korelasi tertinggi (r = 0,93) terhadap popularitas, menegaskan bahwa variabel ini merupakan prediktor paling dominan dalam model klasifikasi. Jumlah rating eksplisit (r = 0,89) dan rata-rata rating (r = 0,74) juga menunjukkan pengaruh yang signifikan, sebagaimana telah ditemukan dalam penelitian Wayesa dkk. yang menempatkan keterlibatan komunitas sebagai prediktor utama popularitas konten buku [6]. Sebaliknya, tahun publikasi memiliki korelasi paling rendah (r = 0,33), mengindikasikan bahwa usia rilis buku bukan determinan utama popularitas pada komunitas Book-Crossing. Perlu ditegaskan bahwa nilai korelasi tinggi antara jumlah pembaca aktif dan kategori popularitas tidak boleh ditafsirkan sebagai bukti kausal yang berdiri sendiri, melainkan harus dipahami dalam konteks bahwa label popularitas dikonstruksikan melalui analisis distribusi jumlah pembaca itu sendiri. Kontribusi substantif dari analisis korelasi ini justru terletak pada temuan bahwa fitur perilaku lain seperti rata-rata *rating*, rentang sebaran *rating*, dan rata-rata usia pembaca memberikan kontribusi prediktif yang signifikan secara independen, sehingga model klasifikasi yang dihasilkan bersifat multivariabel dan tidak semata-mata bergantung pada satu fitur dominan.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 347−356
DOI : 0.47065/bulletinds.v5i3.10449
https://ejurnal.seminar-id.com/index.php/bulletinds

3.4 Evaluasi Performa Model Random Forest Classifier

Model Random Forest Classifier dilatih menggunakan 8.187 data dan diuji pada 2.047 data dengan konfigurasi hyperparameter yang telah ditetapkan. Hasil evaluasi performa model menggunakan empat metrik utama menunjukkan accuracy keseluruhan sebesar 91,2%, dengan nilai F1-Score tertinggi sebesar 0,94 pada kategori Populer, 0,91 pada kategori Cukup Populer, dan 0,89 pada kategori Kurang Populer. Nilai Precision dan Recall pada kategori Populer masing-masing mencapai 0,95 dan 0,93, menandakan bahwa model sangat andal dalam mengidentifikasi buku dengan popularitas tinggi pada dataset Book-Crossing. Performa ini melampaui benchmark yang dilaporkan oleh Wijayanti dan Marcos yang mencapai akurasi 89,19% pada klasifikasi pola peminjaman buku menggunakan Random Forest [8], serta sebanding dengan akurasi 92% yang dilaporkan Roy dan P.D. pada prediksi rating buku berbasis pendekatan hibrida [7]. Temuan ini mengonfirmasi bahwa kombinasi fitur numerik multivariabel berbasis perilaku pembaca, ketika diolah melalui algoritma Random Forest dengan konfigurasi hyperparameter yang dioptimasi melalui grid search, mampu menghasilkan klasifikasi popularitas yang akurat dan dapat diandalkan untuk implementasi pada sistem kurasi konten buku digital secara praktis [4].

3.5 Pola Jumlah Pembaca sebagai Determinan Popularitas Buku Digital

Sebelum membahas pola jumlah pembaca sebagai determinan popularitas, perlu ditegaskan kerangka epistemologis yang melandasi interpretasi hasil dalam sub-bagian ini. Penelitian ini tidak bertujuan untuk membuktikan hubungan sebab-akibat antara jumlah pembaca aktif dan popularitas, melainkan untuk mengevaluasi seberapa konsisten kombinasi fitur perilaku pembaca multivariabel dapat memprediksi kategori popularitas yang telah didefinisikan secara operasional melalui distribusi kuartil pada tahap penetapan label. Dengan kata lain, label popularitas dikonstruksikan terlebih dahulu berdasarkan distribusi empiris jumlah pembaca, kemudian model klasifikasi diuji untuk mengevaluasi apakah kombinasi fitur lain (rata-rata *rating*, rentang sebaran *rating*, rata-rata usia pembaca, tahun publikasi, dan jumlah *rating* eksplisit) mampu mereproduksi kategori tersebut secara akurat. Nilai korelasi *Pearson* sebesar r = 0,93 antara jumlah pembaca aktif dan popularitas, dengan demikian, harus dipahami sebagai konsekuensi logis dari konstruksi variabel target dalam kerangka pengujian model, bukan sebagai klaim temuan kausal yang berdiri sendiri. Kontribusi metodologis penelitian ini justru terletak pada pengujian empiris bahwa fitur perilaku lain memberikan daya prediktif independen yang signifikan, sehingga model klasifikasi yang dihasilkan bersifat multivariabel dan tidak reduktif terhadap satu fitur tunggal.

Berangkat dari kerangka tersebut, hasil analisis secara konsisten menunjukkan bahwa jumlah pembaca aktif merupakan variabel yang paling diskriminatif dalam memisahkan kategori popularitas buku digital pada komunitas *Book-Crossing*. Temuan ini sekaligus membantah asumsi umum yang selama ini menempatkan rata-rata *rating* bintang sebagai satu-satunya tolok ukur kualitas dan popularitas sebuah karya. Jumlah pembaca aktif mencerminkan keterlibatan komunitas secara nyata (*genuine community engagement*) yang tidak mudah dimanipulasi, berbeda dengan sistem *rating* bintang yang rentan terhadap bias subjektivitas individu dan praktik penilaian yang tidak autentik [16]. Kondisi ini diperkuat oleh karakteristik distribusi *rating* pada *dataset Book-Crossing* yang sangat *skewed* ke arah nilai positif, di mana mayoritas *rating* eksplisit berada pada rentang 8-10, sehingga nilai rata-rata *rating* memiliki daya pembeda yang lebih rendah dibandingkan jumlah pembaca aktif dalam mengklasifikasikan popularitas buku secara hierarkis. Artinya, semakin banyak pembaca yang berinteraksi dengan sebuah buku, semakin besar peluang karya tersebut diakui sebagai populer oleh komunitas pembaca yang lebih luas, tanpa bergantung pada penilaian rata-rata yang dapat bias akibat *selection effect* pada pembaca awal.

Menariknya, distribusi popularitas yang ditemukan dalam penelitian ini memperlihatkan pola yang sangat tidak merata: hanya 10,0% buku yang berhasil masuk kategori *Populer*, sementara 58,2% karya lainnya bersaing di segmen *Kurang Populer*. Fenomena ini mencerminkan hukum Pareto dalam ekosistem konten digital, di mana sebagian kecil karya mendominasi sebagian besar perhatian komunitas pembaca, sebagaimana telah dikonfirmasi melalui kajian distribusi *power law* pada berbagai platform konten daring yang menunjukkan bahwa 20% konten teratas mampu menyerap sekitar 80% perhatian publik [17]. Implikasi praktisnya adalah bahwa baik penulis maupun penerbit perlu memahami faktor-faktor struktural yang membedakan buku populer dari yang tidak, bukan sekadar mengandalkan kualitas intrinsik narasi tanpa strategi distribusi yang tepat. Lebih jauh, perlu dipahami bahwa akumulasi jumlah pembaca pada platform komunitas buku tidak terjadi secara acak, melainkan dipengaruhi oleh sejumlah faktor struktural yang saling berkaitan, mulai dari algoritma rekomendasi yang menciptakan efek *snowball*, prestise penulis, hingga ekosistem distribusi fisik buku itu sendiri pada komunitas pertukaran buku internasional. Buku yang telah memiliki basis pembaca awal yang cukup besar cenderung mendapatkan eksposur lebih tinggi melalui mekanisme rekomendasi berbasis *collaborative filtering*, yang pada gilirannya menciptakan siklus yang menguntungkan karya-karya yang sudah dikenal komunitas [18].

Fenomena ini dikenal dalam literatur sebagai *popularity bias* yang inheren pada sistem rekomendasi modern dan cenderung mengaburkan visibilitas karya-karya pada *long tail* distribusi, meskipun karya-karya tersebut mungkin memiliki potensi kualitas yang sebanding [19]. Di samping itu, faktor rata-rata usia pembaca yang memiliki korelasi sedang (r = 0,41) terhadap popularitas juga menunjukkan bahwa segmen demografis tertentu memiliki preferensi yang konsisten terhadap kategori buku tertentu, sebuah informasi yang sangat berharga bagi

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 347−356
DOI : 0.47065/bulletinds.v5i3.10449
https://ejurnal.seminar-id.com/index.php/bulletinds

pengelola platform dalam merancang sistem rekomendasi berbasis demografi yang lebih akurat dan adaptif terhadap karakteristik komunitas pembaca [20]. Dengan demikian, popularitas buku digital bukan semata-mata produk dari kualitas intrinsik karya, melainkan juga hasil dari interaksi kompleks antara prestise penulis, akumulasi *rating* komunitas, dan mekanisme distribusi konten yang berlangsung pada platform pertukaran buku internasional secara berkelanjutan.

3.6 Relevansi Fitur Interaksi dalam Menentukan Daya Tarik Buku

Selain jumlah pembaca aktif, fitur jumlah *rating* eksplisit (r = 0,89) dan rata-rata *rating* (r = 0,74) terbukti memiliki kontribusi signifikan terhadap klasifikasi popularitas buku digital pada *dataset Book-Crossing*. Kedua fitur ini merepresentasikan bentuk keterlibatan aktif (*active engagement*) komunitas pembaca yang melampaui sekadar interaksi pasif. Pembaca yang memberikan *rating* eksplisit pada skala 1-10 menunjukkan komitmen untuk memberikan penilaian terukur terhadap karya yang telah dibaca, sementara akumulasi *rating* dari banyak pembaca mencerminkan validasi sosial yang menghidupkan komunitas pertukaran buku di sekitar sebuah karya. Hasil ini selaras secara substantif dengan temuan yang menyatakan bahwa keterlibatan komunitas pembaca melalui sistem *rating* dan ulasan menjadi prediktor utama keberhasilan sebuah karya jauh melampaui fitur intrinsik buku semata, mengingat *engagement* aktif mampu menciptakan efek difusi sosial yang memperluas jangkauan karya secara eksponensial [19].

Kajian lebih mendalam terhadap pola *rating* pada buku-buku yang masuk kategori *Populer* dalam *dataset* penelitian ini mengungkap fakta menarik: lebih dari 60% *rating* eksplisit yang masuk pada buku populer berada pada skala 8-10, yakni pembaca mengekspresikan apresiasi tinggi terhadap karya tersebut. Pola ini berbeda secara signifikan dengan buku kategori *Kurang Populer* yang cenderung menerima *rating* lebih sedikit dengan distribusi yang lebih merata di rentang 5-7. Temuan ini menunjukkan bahwa *rating* eksplisit bukan hanya berfungsi sebagai indikator popularitas pasif, melainkan juga sebagai mekanisme *quality signaling* aktif yang secara psikologis memotivasi pembaca baru untuk turut mengeksplorasi karya tersebut. Fenomena ini memiliki implikasi yang sangat relevan bagi pengelola platform komunitas buku. Alih-alih hanya menyediakan kolom *rating* sebagai fitur standar, platform perlu merancang sistem gamifikasi (*gamification*) yang mendorong pembaca untuk lebih aktif memberikan penilaian eksplisit, misalnya melalui pemberian lencana (*badge*) bagi pembaca paling aktif, sistem *streak* harian, atau fitur *reaction* yang lebih beragam dibandingkan sekadar pemberian skor numerik. Strategi semacam ini telah terbukti efektif dalam meningkatkan retensi pengguna pada berbagai platform konten digital global, di mana penerapan elemen gamifikasi mampu menurunkan tingkat *churn* secara signifikan dan memperkuat ikatan emosional pengguna terhadap platform [21].

Penerapan strategi serupa pada platform pertukaran buku diyakini dapat menciptakan ekosistem yang lebih dinamis dan kompetitif. Tidak kalah penting, fenomena rentang sebaran *rating* (r = 0,62) sebagai prediktor popularitas juga perlu dimaknai secara lebih kontekstual. Data penelitian ini menunjukkan bahwa buku dengan rentang sebaran *rating* yang lebar tidak selalu memiliki jumlah pembaca yang sebanding, yang mengindikasikan adanya segmen pembaca dengan preferensi yang sangat *polarized* terhadap karya tertentu. Segmen pembaca ini justru berpotensi menjadi aset tersembunyi bagi popularitas sebuah karya, karena mereka cenderung memicu diskusi komunitas yang lebih intens, baik dalam bentuk pujian maupun kritik konstruktif. Oleh karena itu, metrik sebaran *rating* layak mendapatkan bobot yang lebih signifikan dalam algoritma rekomendasi platform sebagai cerminan minat komunitas yang lebih kompleks daripada sekadar nilai rata-rata. Perspektif ini sejalan dengan kajian dinamika *feedback loop* pada sistem rekomendasi yang menunjukkan bahwa interaksi berulang antara algoritma dan pengguna mampu memperkuat atau melemahkan visibilitas sebuah karya secara akumulatif, tergantung pada bagaimana metrik keterlibatan ditangkap dan dimodelkan [22].

3.7 Pembahasan

Apabila dibandingkan dengan penelitian-penelitian terdahulu, terdapat beberapa perbedaan mendasar yang menegaskan kebaruan kontribusi penelitian ini. Penelitian terdahulu yang menggunakan *dataset Book-Crossing* lebih banyak berorientasi pada prediksi *rating* individual atau personalisasi rekomendasi melalui pendekatan *deep learning* berbasis *Singular Value Decomposition* (SVD), tanpa menempatkan klasifikasi popularitas multikelas sebagai tujuan utama [23]. Penelitian ini justru membalik paradigma tersebut dengan menempatkan jumlah pembaca aktif sebagai prediktor primer dan membuktikan efektivitasnya secara empiris melalui nilai korelasi yang sangat tinggi (r = 0,93) serta akurasi klasifikasi multikelas mencapai 91,2%. Aspek komparasi dengan penelitian terdahulu juga perlu diperluas pada dimensi metodologis. Sebagian besar penelitian sebelumnya yang mengkaji popularitas buku digital pada *dataset Book-Crossing* lebih memilih pendekatan *collaborative filtering* murni yang sangat bergantung pada matriks interaksi *user-item* berskala besar. Pendekatan ini rentan terhadap masalah *cold start* dan *data sparsity*, terutama mengingat tingkat *sparsity* pada *dataset Book-Crossing* yang mencapai 0,0015% menurut analisis komparatif lintas *dataset* publik [24].

Penelitian ini mengambil langkah berbeda dengan menggunakan pendekatan klasifikasi multikelas berbasis fitur agregat per-ISBN, sehingga hasil yang diperoleh lebih *robust* terhadap masalah *cold start* karena tidak bergantung pada matriks interaksi *user-item* yang penuh dengan nilai kosong. Dari perspektif komparasi performa model, akurasi 91,2% yang dihasilkan oleh *Random Forest Classifier* dalam penelitian ini melampaui beberapa *benchmark* yang dilaporkan dalam literatur sejenis dengan pendekatan *ensemble learning* hibrida. Perbedaan ini

ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 347−356
DOI : 0.47065/bulletinds.v5i3.10449
https://ejurnal.seminar-id.com/index.php/bulletinds

mengisyaratkan bahwa kombinasi fitur perilaku pembaca berskala besar dengan algoritma *ensemble* yang dioptimasi melalui *grid search* memberikan daya prediktif yang lebih superior dalam konteks klasifikasi popularitas buku digital. Meskipun demikian, perlu diakui bahwa penelitian ini memiliki keterbatasan dalam hal periode pengumpulan data, mengingat *dataset Book-Crossing* dihimpun pada tahun 2004 sehingga belum merepresentasikan dinamika perilaku pembaca pasca-revolusi *e-reader* dan *audiobook* yang berkembang pesat. Penelitian lanjutan yang mengintegrasikan data dari platform yang lebih kontemporer dengan metode *transfer learning* diharapkan dapat menghasilkan model yang lebih generalisabel dan adaptif terhadap perilaku pembaca digital masa kini. Dengan demikian, kontribusi penelitian ini hendaknya dipandang sebagai fondasi awal yang membuka ruang eksplorasi lebih lanjut dalam bidang analitik perbukuan digital berbasis data, sekaligus menggarisbawahi pentingnya pendekatan *fairness-aware machine learning* dalam pemodelan popularitas buku agar sistem yang dibangun tidak hanya akurat secara teknis, tetapi juga berkeadilan bagi seluruh ekosistem penulis dan pembaca tanpa memandang skala atau prestise.

4. KESIMPULAN

Penelitian ini secara komprehensif telah berhasil menganalisis kombinasi fitur perilaku pembaca sebagai prediktor dalam menentukan kategori popularitas karya pada *dataset Book-Crossing* yang dihimpun dari komunitas pertukaran buku internasional. Melalui serangkaian tahapan sistematis mulai dari pengumpulan data sekunder berbasis platform *Kaggle*, *preprocessing*, ekstraksi fitur, pemodelan *machine learning*, hingga evaluasi model, dengan tiga temuan substantif utama yaitu terkonfirmasinya peran jumlah pembaca aktif, dengan nilai korelasi *Pearson* sebesar r = 0,93 yang harus dimaknai dalam kerangka konstruksi label popularitas dari distribusi data dan bukan sebagai klaim kausal, efektivitas algoritma *Random Forest Classifier* dengan *accuracy* 91,2% dan *F1-Score* tertinggi 0,94 pada kategori *Populer* yang membuktikan superioritas integrasi fitur numerik multivariabel berbasis perilaku pembaca, serta distribusi popularitas yang sangat tidak merata di mana hanya 10,0% buku masuk kategori *Populer* dan 58,2% berada di kategori *Kurang Populer* mencerminkan fenomena *power law* sekaligus *popularity bias* yang inheren pada ekosistem konten digital, sehingga secara teoretis penelitian ini berkontribusi pada penajaman kerangka konseptual pengukuran popularitas sebagai konstruk multidimensi yang dibedakan secara eksplisit dari *rating* individu dan diuji sebagai variabel kategorikal hasil konstruksi distribusi empiris, sementara secara praktis hasil penelitian memberikan tiga implikasi konkret yaitu bagi penulis dan penerbit untuk mengarahkan strategi promosi pada peningkatan keterlibatan komunitas pembaca secara luas bukan semata pada pencapaian *rating* tinggi dari pembaca terbatas, bagi pengelola platform digital untuk merancang sistem rekomendasi yang memperhitungkan *popularity bias* agar karya pada *long tail* distribusi tetap memiliki visibilitas yang adil melalui pendekatan *fairness-aware machine learning*, dan bagi pengembang sistem klasifikasi konten untuk memanfaatkan kombinasi fitur perilaku pembaca multivariabel sebagai pendekatan yang lebih informatif dibandingkan indikator tunggal konvensional, dengan keterbatasan penelitian berupa periode pengumpulan data tahun 2004 yang belum merepresentasikan dinamika perilaku pembaca pasca-revolusi *e-reader*, serta belum terintegrasinya analisis sentimen dari ulasan tekstual. Penelitian berikutnya disarankan untuk memperluas cakupan dengan integrasi data platform kontemporer melalui pendekatan *transfer learning* dan *fairness-aware machine learning* guna menghasilkan model prediksi popularitas yang lebih generalisabel dan berkeadilan bagi seluruh ekosistem penulis maupun pembaca.

REFERENCES

[1] M. Zhou, G. H. Chen, P. Ferreira, and M. D. Smith, “Consumer Behavior in the Online Classroom: Using Video Analytics and Machine Learning to Understand the Consumption of Video Courseware,” *J. Mark. Res.*, vol. 58, no. 6, pp. 1079–1100, 2021, doi: 10.1177/00222437211042013.

[2] S. Daniil, M. Cuper, C. C. S. Liem, J. van Ossenbruggen, and L. Hollink, *Hidden Author Bias in Book Recommendation*, vol. 1, no. 1. Association for Computing Machinery, 2022. [Online]. Available: http://arxiv.org/abs/2209.00371

[3] S. Tripathi, A. Tiwari, K. Upreti, and G. V. Radhakrishnan, “A Machine Learning Approach to Consumer Behavior Analysis in Social Media-Influenced E-Book Markets,” *J. Appl. Sci. Technol. Trends*, vol. 6, no. 2, pp. 242–250, 2025, doi: 10.38094/jastt62318.

[4] E. Abdu and A. Mamuye, “Book Recommendation Using Collaborative Filtering Algorithm,” *Appl. Comput. Intell. Soft Comput.*, vol. 2023, pp. 1–12, Mar. 2023, doi: 10.1155/2023/1514801.

[5] F. Wayesa, M. Leranso, G. Asefa, and A. Kedir, “Pattern-based hybrid book recommendation system using semantic relationships,” *Sci. Rep.*, vol. 13, no. 1, pp. 1–12, 2023, doi: 10.1038/s41598-023-30987-0.

[6] Y. Afoudi, M. Lazaar, and M. Al Achhab, “Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network,” *Simul. Model. Pract. Theory*, vol. 113, p. 102375, 2021, doi: https://doi.org/10.1016/j.simpat.2021.102375.

[7] Ega Ranaldi Pebriansyah, Susanti, Rahmiati, and Triyani Arita Fitri, “Prediction of Library Book Borrowing Patterns Using The Random Forest Algorithm,” *J. Ris. Inform.*, vol. 7, no. 4, pp. 327–335, 2025, doi: 10.34288/jri.v7i4.409.

[8] M. Naghiaei, H. A. Rahmani, and Y. Deldjoo, “CPFair: Personalized Consumer and Producer Fairness Re-ranking for Recommender Systems,” in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 770–779. doi: 10.1145/3477495.3531959.

[9] P. K. Sahoo, V. R. S. Dhanish, and A. N. Kumar, “Book Recommendation Using Collaborative Filtering Algorithm,” *Appl. Comput. Intell. Soft Comput.*, vol. 2023, no. 04, pp. 365–369, 2023, doi: 10.1155/2023/1514801.

[10] Y. Du, L. Peng, S. Dou, X. Su, and X. Ren, “Research on Personalized Book Recommendation Based on Improved Similarity Calculation

and Data Filling Collaborative Filtering Algorithm," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Sep. 2022, doi: 10.1155/2022/1900209.

[11] C. J. Arizmendi *et al.*, "Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work," *Behav. Res. Methods*, vol. 55, no. 6, pp. 3026–3054, 2023, doi: 10.3758/s13428-022-01939-9.

[12] H. Wang and N. Li, "A Click-Through Rate Prediction Method Based on Cross-Importance of Multi-Order Features," 2024, [Online]. Available: http://arxiv.org/abs/2405.08852

[13] M. Cinelli, G. de Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, "The echo chamber effect on social media," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, no. 9, 2021, doi: 10.1073/pnas.2023301118.

[14] J. Prent and M. Mansoury, "Correcting Popularity Bias in Recommender Systems via Item Loss Equalization," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:273186697

[15] A. Klimashevskaia, D. Jannach, M. Elahi, and C. Trattner, "A survey on popularity bias in recommender systems," *User Model. User-adapt. Interact.*, vol. 34, no. 5, pp. 1777–1834, 2024, doi: 10.1007/s11257-024-09406-0.

[16] M. Mansoury, F. Duijvestijn, and I. Mourabet, "Potential Factors Leading to Popularity Unfairness in Recommender Systems: A User-Centered Analysis," pp. 1–16, 2023, [Online]. Available: http://arxiv.org/abs/2310.02961

[17] Y. Zoralioglu and E. Yalcin, "Dynamic feedback loops in recommender systems: Analyzing fairness, popularity bias, and user group disparities," *J. Intell. Inf. Syst.*, 2026, doi: 10.1007/s10844-026-01025-y.

[18] X. Yang *et al.*, "Click-through rate prediction using transfer learning with fine-tuned parameters," *Inf. Sci. (Ny).*, vol. 612, pp. 188–200, 2022, doi: https://doi.org/10.1016/j.ins.2022.08.009.

[19] K. Xiang, *Research on Deep Learning Recommendation System Based on Book Scoring*, *jurnal inf. Sci* 2025. doi: 10.1145/3718751.3718838.

[20] Pratama, "Analisis Sentimen BRImo dan BCA Mobile Menggunakan Support Vector Machine dan Lexicon Based," *Jurnal Ilmu Komputer dan Informatika.*, vol. 3, no. 2, pp. 1439-1450, 2023.

[21] Damayanti *et al.*, "Sentiment Analysis of Alfagift Application User Reviews Using Long Short-Term Memory (LSTM) and Support Vector Machine (SVM) Methods," *Jurnal DECODE*, vol. 4, no. 2, pp. 509-521, 2024, doi: 10.51454/decode.v4i2.478.

[22] Arumi,, *Implementation of Naïve bayes Method for Predictor Prevalence Level for Malnutrition Toddlers in Magelang City*. *Jurnal RESTI.*,2023, doi: 10.29207/resti.v7i2.4438

[23] Jayaditya, I. K. A., & Kadyanan, I. G. A. G. A., Implementasi Random Forest pada Klasifikasi Penyakit Kardiovaskular dengan Hyperparameter Tuning Grid Search. *Jurnal Nasional Teknologi Informasi dan Aplikasinya (JNATIA),* 2023, doi: 10.24843/JNATIA.2023.v02.i01.p25.

[24] Sholihah, N., & Hermawan, A., Implementation of Random Forest and SMOTE Methods for Economic Status Classification in Cirebon City. *Jurnal Teknik Informatika (JUTIF),* 2023, doi: 10.52436/1.jutif.2023.4.6.1135.