JURNAL ILMIAH EKONOMI DAN BISNIS, Vol.13, No.2, Desember 2020, pp. 89 - 96
p-ISSN : 1979-0155 (print)
e-ISSN : 2614-8870 (online)
http://journal.stekom.ac.id/index.php/E-Bisnis

Algoritma Klasifikasi *Decision Tree* Untuk Rekomendasi Buku Berdasarkan Kategori Buku

Mawadatul Maulidah[1], Windu Gata[2], Rizki Aulianita[3], Cucu Ika Agustyaningrum[4]
Program Studi Ilmu Komputer, STMIK Nusa Mandiri[1,2,4]
Program Studi Sistem Informasi, STMIK Nusa Mandiri[3]
Jalan Kramat Raya No. 18, Senen, Jakarta Pusat, (021) 31908575
Email : 14002373@nusamandiri.ac.id[1], windu@nusamandiri.ac.id[2], rizki.rzk@nusamandiri.ac.id[3], 14002365@nusamandiri.ac.id[4]

ARTICLE INFO

Article history:

Received 30 September 2020
Received in revised form 2 Oktober 2020
Accepted 10 Oktober 2020
Available *online* 22 Oktober 2020

ABSTRACT

With the increasing development of technology the more variety of books circulating on the internet. As is the recommendation system on online book sites that provide books relevantly and as needed with one's preferences. One alternative is GoodReads, a social networking site that specializes in cataloging books and users can share reading book recommendations with each other by rating, reviewing, and commenting. As a large book recommendation site, it has a lot of data that can be processed by applying machine learning methods, but still not known as the most accurate model. By using the right model, we can provide more accurate recommendations. Therefore, this study will analyze the data obtained from the www.kaggle.com namely the goodreads-books dataset. This study proposed a data mining classification model to get the best model in recommending books on GoodReads. The algorithms used are Decision Tree, K-Nearest Neighbor, Naïve Bayes, Random Forest, and Support Vector Classifier, then for model evaluation using accuracy, precision, recall, f1-score, confusion matrix, AUC, and Mean Error Absolute. The test results of several classification algorithms found that Decision Tree has the highest accuracy among the methods presented by 99.95%, precision by 100%, recall by 96%, f1-score of 98% with MAE of 0.05 and AUC of 99.96%. This is proof that decision tree algorithms can be used as book recommendations based on book categories on GoodReads.

Keywords: data mining, decision tree, GoodReads, classification.

1. PENDAHULUAN

Meningkatnya perkembangan teknologi yang semakin cepat membuat semakin banyak ragam buku yang beredar di internet. Dengan adanya sistem rekomendasi situs buku *online* dapat menyediakan buku yang relevan dan dibutuhkan sesuai dengan selera seseorang, hampir semua situs buku *online* sudah menggunakan sistem rekomendasi. Beberapa contoh situs yang telah menerapkan metode sistem rekomendasi seperti *Goodreads*, *Google*, *Openlibrary*, *Scribd* dan masih banyak lainnya. Pada dasarnya, situs buku *online* yang

Received September 30, 2020; Revised Oktober 2, 2020; Accepted Oktober 22, 2020

p-ISSN : 1979-0155 e-ISSN : 2614-8870

sudah menerapkan sistem rekomendasi namun tidak selalu menghasilkan rekomendasi yang diharapkan oleh pengguna. Hal ini disebabkan karena sistem rekomendasi memprediksi rekomendasi suatu produk yang memiliki rating tinggi atau berdasarkan riwayat orang yang mengakses kategori buku tertentu pada situs *online*. Sistem rekomendasi yang tidak memperhatikan seluruh preferensi pengguna dapat mengakibatkan rekomendasi yang tidak tepat dan tidak akurat. Salah satu dari situs terbesar di dunia yang fokus pada pembaca dan rekomendasi buku yaitu *goodreads*.com.

Goodreads merupakan komunitas baca internasional yang digagas Otis Chandler pada tahun 2006 dan diluncurkan pada 30 Januari 2007 bersama Elizabeth Khuri. Pada *Goodreads* penggunanya dapat memberikan peringkat dan menulis ulasan tentang buku serta berdiskusi dengan pengguna lainnya. Pada *GoodReads* pengguna dapat saling berbagi rekomendasi buku bacaan dengan memberikan rating, *review* maupun komentar [1]. Daya tarik *Goodreads* bagi pengguna pada dasarnya ada tiga. Pertama, *Goodreads* memfasilitasi pengkatalogisasi buku berdasarkan daftar buku sudah dibaca (*read*), buku yang sedang dibaca (*currently reading*), dan akan dibaca (*to read*). Kedua, situs ini menyediakan ulasan dan penilaian buku (baik pribadi maupun kumulatif), memfasilitasi transisi pembaca dari yang pasif menjadi kritikus amatir. Ketiga, *Goodreads* memberikan rekomendasi buku yang disesuaikan secara individual yang dihasilkan melalui pemfilteran kolaboratif dari preferensi pembaca lain [2]. Selain ulasan, *Goodreads* juga memberikan fasilitas rekomendasi dan rating yang dapat membantu pembaca untuk memilih buku yang relevan. Namun terkadang pengguna *Goodreads* hanya membaca tanpa memberikan rating sehingga pengikut dari pengguna tersebut ingin tahu berapa rating yang diberikan oleh pengguna tersebut pada buku tertentu[2].

Data yang tersedia pada *Goodreads* API untuk publik saat ini semakin banyak, sehingga kesempatan menerapkan metode machine learning untuk prediksi maupun sistem rekomendasi juga semakin banyak digunakan. Saat ini para peneliti telah mengembangkan berbagai metode untuk membangun pemodelan pada data *Goodreads* untuk memprediksi rating buku atau sebagai sistem rekomendasi dengan menggunkan data mining. Data mining merupakan proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu [3]. Klasifikasi adalah salah satu model dalam data mining. Model klasifikasi merupakan teknik memprediksi data, membuat prediksi nilai dari suatu data yang hasilnya telah ditemukan berasal dari data yang berbeda. Tujuan dari model ini yaitu memprediksi nilai dari suatu variabel yang tidak diketahui dari variabel lain yang telah diberikan [4]. Metode yang dapat digunakan dalam prediksi adalah metode klasifikasi seperti *Decision Tree, Random Forest, Naïve Bayes, KNN, ANN, SVM, Regresi* dan algoritma lainnya. *Decision Tree* disebut juga dengan pohon keputusan merupakan model dari klasifikasi. Bentuknya yang seperti struktur pohon merepresentasikan atribut setiap data yang diproses. *Decision Tree* yaitu mendeskripsikan tiap-tiap kelas untuk menemukan pola atau fungsi dengan tujuan melakukan klasifikasi atau prediksi data yang belum mempunyai kelas. Metode ini sangat populer karena dapat dipakai pada banyak bidang. Beberapa algoritma yang sering yang menerapkan *Decision Tree* yaitu C4.5, ID3, dan *Random Forest* [4]. Salah satu contoh penerapan metode yang digunakan oleh dataset *goodreads books* adalah metode *Artificial Neural Network* (ANN) untuk Memprediksi Peringkat Keseluruhan Buku [5]. Pada penelitian yang dilakukan oleh [5] Prediksi didasarkan pada beberapa fitur *(bookID, title, authors, isbn, language_code, isbn13, # num_pages, ratings_count, text_reviews_count),* yang digunakan sebagai variabel masukan dan sebagai variabel keluaran untuk model prediksi ANN yaitu (*average_rating*). Selain itu, kumpulan data tersebut berisi 13720 instance. Setelah pra-pemrosesan, ini menjadi 12241 yang merupakan jumlah yang besar untuk ditangani oleh jaringan neural, jadi data menjadi 8242 instance pelatihan, dan 3999 instance validasi. Menghasilkan nilai prediksi sebesar 99,78%.

Penelitian terkait yang sudah dilakukan untuk prediksi rating buku dengan dataset yang sama diantaranya yaitu Implementasi Algoritma *Naïve Bayes* Dalam Penentuan Rating Buku oleh Rizki Ayuning Tyas *et al* [6] dengan data testing sebanyak 2226 memperoleh hasil penentuan rating buku dengan accuracy 66,98%, precision 74,47% dan recall 62,47% dan hasil analisis ini di dapatkan dari dataset yang ada pada situs www.kaggle.com menunjukan bahwa mayoritas prediksi rating buku cenderung rendah. Penelitian sebelumnya dalam menggunakan Metode *Decision Tree* oleh Afrilio Franseda [4] dengan judul Integrasi Metode *Decision Tree* dan SMOTE untuk Klasifikasi Data Kecelakaan Lalu Lintas. Dalam penelitian ini proses untuk meningkatkan evaluasi menggunakan proses Knowledge Discovery in Database (KDD). Pengujian menggunakan Split Validation *Decision Tree* dan SMOTE diperoleh akurasi 69.23%. Pengujian pada Cross Validation *Decision Tree* dan SMOTE diperoleh akurasi 63.56%. Pada pengujian *Decision Tree* dan SMOTE Split Data diperoleh akurasi 71.12% dengan perbandingan 1:9. Hasil dari perbandingan diperoleh bahwa *Decision Tree* dan SMOTE Split Data mendapatkan akurasi yang paling baik dengan presisi 89.71% (3:7) dan area under curve (AUC) sebesar 0.773 (1:9). Selanjutnya penelitian sebelumnya terkait Komparasi Algoritma C4.5 Dan *Naive Bayes* Yang Dikembangkan Menjadi Web Intellegence Pada

p-ISSN : 1979-0155 e-ISSN : 2614-8870■

Perhitungan Bonus Tahunan Karyawan Di PT. Abc oleh Taransa Agasya Tutupoly dan Ibnu Alfarobi [7] pada penelitian ini menggunakan pembagian data testing : data training 10 : 90, 20 : 80, 30 : 70. Perbandingan dengan nilai akurasi algoritma *Naive Bayes* pada percobaan data testing 10% : data training 90% mendapat nilai akurasi yang terbesar sedangkan nilai akurasi dengan menggunakan algoritma klasifikasi C4.5 yang terbesar pada percobaan data testing 30% : data training 70%. Sedangkan evaluasi menggunakan ROC curve yaitu berdasarkan nilai AUC, algoritma *Naive Bayes* menjadi yang tertinggi pada percobaan data testing 20% : data training 80% memperoleh nilai 0.990 sedangkan untuk data testing 30% : data training 70% dengan nilai 0.991. Berdasarkan hasil dari keseluruhan pengujian model algoritma dapat disimpulkan bahwa kinerja C4.5 dan *Naïve Bayes* hampir sama bagusnya, baik itu dilihat dari tingkat akurasi maupun AUC nya.

Oleh karena itu dari uraian diatas, berbagai metode yang telah digunakan untuk dapat merekomendasi buku dari prediksi kategori buku apa yang layak dibaca dan diminati oleh pengguna ini masih belum menunjukkan hasil yang akurat, karena dari beberapa metode memiliki akurasi yang masih kurang baik. Penelitian ini bertujuan untuk mendapatkan model algoritma terbaik untuk rekomendasi buku berdasarkan prediksi kategori buku pada dataset *goodreads books* dengan menggunakan beberapa metode yang ada pada data mining berdasarkan faktor-faktor yang mempengaruhi rekomendasi buku dari prediksi katergori buku dengan melakukan pengujian nilai *accuracy, precision, recall, f1-score, confusion matrix, AUC* dan *Mean Absolute Error.* Metode klasifikasi data mining yang digunakan yaitu *Decision Tree, K-Nearest Neighbor (K-NN), Naïve Bayes, Random Forest* dan *Support Vector Classifier (SVC).* Hasil dari penelitian ini diharapkan bermanfaat untuk pengguna *Goodreads* untuk dapat merekomendasikan buku yang layak dan banyak diminati.

2. METODOLOGI PENELITIAN

2.1. Dataset

Dalam penelitian ini untuk rekomendasi buku dari prediksi kategori buku menggunakan sampel dari dataset *goodreads books* yang dibuat oleh pengguna Soumik [8] dari situs *www.kaggle.com*. Data yang didapatkan adalah data berupa data statistik buku pada *website www.goodreads.com*[1]. Data yang terkumpul sebanyak 11.127 data yang terdiri dari atribut *bookID, title. author, average_rating, isbn, isbn13, language_code, num_pages, rating_count, text_review _count, publication_date, publisher*. Berikut informasi dari dataset *goodreads books*:

Tabel 1. Informasi dataset *goodreads*[8]

#	Atribut	Deskripsi	Tipe
1.	*bookID*	Nomor Identifikasi unik untuk setiap buku.	*integer*
2.	*tittle*	Nama buku yang diterbitkan.	*string*
3.	*authors*	Nama penulis buku. Beberapa penulis dipisahkan dengan -.	*string*
4.	*average_rating*	Nilai rata-rata dari buku yang diterima secara total.	*real*
5.	*isbn*	Nomor unik lain untuk mengidentifikasi buku tersebut, Nomor Buku Standar Internasional.	*long*
6.	*language_code*	Membantu memahami apa bahasa utama buku tersebut. Misalnya, eng adalah standar bahasa Inggris.	*string*
7.	*isbn13*	ISBN 13-digit untuk mengidentifikasi buku, bukan ISBN standar 11-digit.	*long*
8.	*# num_pages*	Jumlah halaman buku tersebut.	*integer*
9.	*rating_count*	Jumlah total peringkat yang diterima buku.	*integer*
10	*text_review_count*	Jumlah total ulasan teks tertulis yang diterima buku.	*integer*
11	*publication_date*	Tanggal buku yang diterbitkan	*string*
12	*publisher*	Nama tempat buku diterbitkan	*string*

2.2. Tahapan Penelitian

Tahapan Penelitian yang digunakan dalam penelitian ini menggunakan metode Algoritma *Decision Tree* yang akan dibandingkan dengan beberapa algoritma klasifikasi seperti *Random Forest*, SVC, KNN dan *Naïve Bayes* dengan menggunakan bantuan *tools* berupa bahasa pemrograman *Pyhton Jupiter* untuk mencari algoritma terbaik sebagai prediksi untuk menentukan rekomendasi buku mana yang banyak diminati oleh pengguna. Langkah-langkah yang dilakukan dalam pengolahan data, sebagai berikut:

2.2.1. Tahapan *Selection*

p-ISSN : 1979-0155 e-ISSN : 2614-8870

Tahapan ini bertujuan untuk pemilihan dataset yang akan digunakan dalam penelitian ini. Data yang dipilih harus sesuai dengan batasan yang sudah ditentukan di dalam penelitian ini berupa data yang digunakan data selama beberapa tahun dimulai dari 1900–2019. Dalam tahapan ini dataset yang dimiliki akan dibagi menjadi data training dan data testing.

2.2.2. Tahapan *Pre-Processing*

Tahapan *pre-processing* untuk mencari data yang memiliki missing value yang berarti data yang didapat belum normal. Namun untuk dataset *goodreads-books* tidak terdapat missing value. Kemudian didalam tahapan pre-processing ini terdapat tahapan membuat label/*Class*, merubah atribut String menjadi Numerik.

a. Membuat class dengan label *"Best Books", "Good Books","Not-Bad Books"* dan *"Lowest Rated Books"* dari atribut *average_rating*.
b. Proses String ke Numerik mengubah tipe atribut string menjadi tipe numerik. Operator ini tidak hanya mengubah tipe atribut yang dipilih tetapi juga memetakan semua nilai atribut ini ke nilai numerik yang sesuai.

2.2.3. Tahapan Pemilihan Model

Algoritma Decision Tree berasal dari algoritma *Concept Learning System (CLS)* dan *Iterative Dichotomiser 3 (ID3)*. *Decision Tree* berdasarkan algoritma C4.5 adalah teknik klasifikasi yang umum digunakan untuk mengekstrak hubungan yang relevan dalam data. Algoritma C4.5 adalah program yang membuat pohon keputusan berdasarkan pada set data input berlabel. Kelebihannya adalah modelnya dapat dengan mudah ditafsirkan dan diimplementasikan dengan nilai kontinu dan nilai diskrit. Algoritma C4.5 membagi data training dengan bantuan perolehan informasi. Atribut yang memiliki frekuensi tinggi dipertimbangkan untuk memisahkan data berdasarkan informasi yang tersedia dalam dataset.

Sebelum menghitung nilai gain terlebuh dahulu untuk mengetahui nilai entropy yaitu dengan persamaan untuk sebagai berikut:

$$Entropy(S) = \sum_{i=1}^{n} pi * log_2\ pi \quad (1)$$

Keterangan:

S artinya Himpunan kasus
A artinya Atribut
n artinya Jumlah partisi S
pi artinya Proporsi dari Si terhadap S

Persamaan yang digunakan untuk menghitung *Information Gain*:

$$Gain(\mathrm{S}, \mathrm{A}) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (2)$$

Keterangan:

S artinya Himpunan kasus
A artinya Atribut
n artinya Jumlah partisi atribut
A |Si| artinya Jumlah kasus pada partisi ke-i
|S| artinya Jumlah kasus dalam S

Secara ringkas, tahapan algoritma *Decision Tree* dapat digambarkan sebagai berikut:

a. Menghitung nilai *Information Gain* dari setiap atribut.
b. Memilih atribut yang memiliki nilai *Information Gain* paling besar.
c. Membentuk simpul yang berisi atribut tersebut.
d. Proses perhitungan *Information Gain* akan terus dilaksanakan sampai semua data telah termasuk dalam kelas yang sama. Atribut yang telah dipilih tidak diikutkan lagi dalam perhitungan nilai *Information Gain* selanjutnya.

Algoritma yang digunakan dalam pengujian ini yaitu *Decision Tree* sebagai algoritma dasar yang akan di komparasikan dengan algoritma *Random Forest, KNN, SVC dan Naïve Bayes.*

2.2.4. Tahapan Evaluasi dan Validasi Hasil

Tahapan ini digunakan untuk memperoleh prediksi menggunakan model algoritma yang sudah ditentukan kemudian membandingkannya dengan beberapa hasil model algoritma yang lainnya. Tahapan ini mewakili langkah penting dalam proses membangun sebuah model. Tahapan ini juga melakukan Percentage split_train_test sebesar (80%), salah satu model yang akan digunakan pada penelitian ini yaitu mengevaluasi hasil akurasi dari algortima dengan menggunakan 80% dari dataset sebanyak 8684 data sebagai data training

dan 20% dari dataset sebanyak 2171 data sebagai data testing. Untuk pengujian menggunakan nilai *accuracy*, *precision*, *recall*, *f1-score*, *Mean Absolute Error*, *Confusion matrix* dan nilai ROC/AUC. Untuk membuktikan kinerja algoritma yang digunakan menggunakan persamaan sebagai berikut:

$$Accurancy = \frac{TP + TN}{TP + FP + TN + FN} x100\% \quad (3)$$

Akurasi yang dihasilkan dihitung berdasarkan *confusion matrix*. Perhitungan pada confusion matrix dihitung sesuai dengan prediksi positif yang benar (*True Positif*), prediksi positif yang salah (*False Positif*), prediksi negatif yang benar (*True Negatif*) dan prediksi negatif yang salah (*False Negatif*)[9]. Semakin tinggi nilai akurasi yang didapat maka semakin baik pula metode yang dihasilkan.

Precision adalah rasio prediksi benar positif (TP) yang dibandingkan dengan keseluruhan hasil yang diprediksi positif.

$$Precision = \frac{TP}{TP + FP}\ x\ 100\% \quad (4)$$

Recall adalah rasio prediksi benar positif (TP) yang dibandingkan dengan keseluruhan data yang benar positif.

$$Recall = \frac{TP}{TP + FN} x\ 100\% \quad (5)$$

F1 Score adalah perbandingan rata-rata presisi dan recall yang dibobotkan.

$$F1 - score = 2x \frac{Precision\ x\ Recall}{(Precision + Recall)} \quad (6)$$

Confussion matrix adalah alat (tools) visualisasi yang biasa digunakan untuk menganalisis seberapa baik kualitas pengklasifikasi dapat mengenali data dari kelas yang berbeda. Sedangkan ROC menurut adalah ukuran numerik untuk membedakan kinerja model, dan menunjukkan seberapa sukses dan benar peringkat model dengan memisahkan pengamatan positif dan negatif[10]. Klasifikasi akurasi menggunakan *area under the curve* (AUC) dinyatakan dalam lima kategori, yaitu

Akurasi bernilai 0.90 – 1.00 = *Excellent classification*
Akurasi bernilai 0.80 – 0.90 = *Good classification*
Akurasi bernilai 0.70 – 0.80 = *Fair classification*
Akurasi bernilai 0.60 – 0.70 = *Poor classification*
Akurasi bernilai 0.50 – 0.60 = *Failure excellent*[11].

Pada tahapan ini akan ditampilkan dalam bentuk table dan grafik untuk komparasi algoritma klasifikasinya.

3. HASIL DAN PEMBAHASAN

Berdasarkan penelitian yang dilakukan pada dataset *goodreads-books* dengan melakukan *Percentage split_train_test* (80%), salah satu model yang akan digunakan pada penelitian ini yaitu mengevaluasi hasil akurasi dari algortima dengan menggunakan 80% dari dataset sebanyak 8684 data sebagai data *training* dan 20% dari dataset sebanyak 2171 data sebagai data *testing*. Dataset ini memiliki 12 atribut sebagai variabel masukan dan *RatingCategory* sebagai variabel keluaran yang dibuat dari (*average_rating*) dengan praproccesing yang mempunyai 4 kelas yaitu *Best Books*, *Good Books*, *Not-Bad Books dan Lowest Rated Books*. Hasil penelitian akan dijelaskan serta dapat dilihat lebih rincinya pada tabel 2. dan akan ditampilkan dalam bentuk grafik histogram pada gambar 2., sebagai berikut ini:

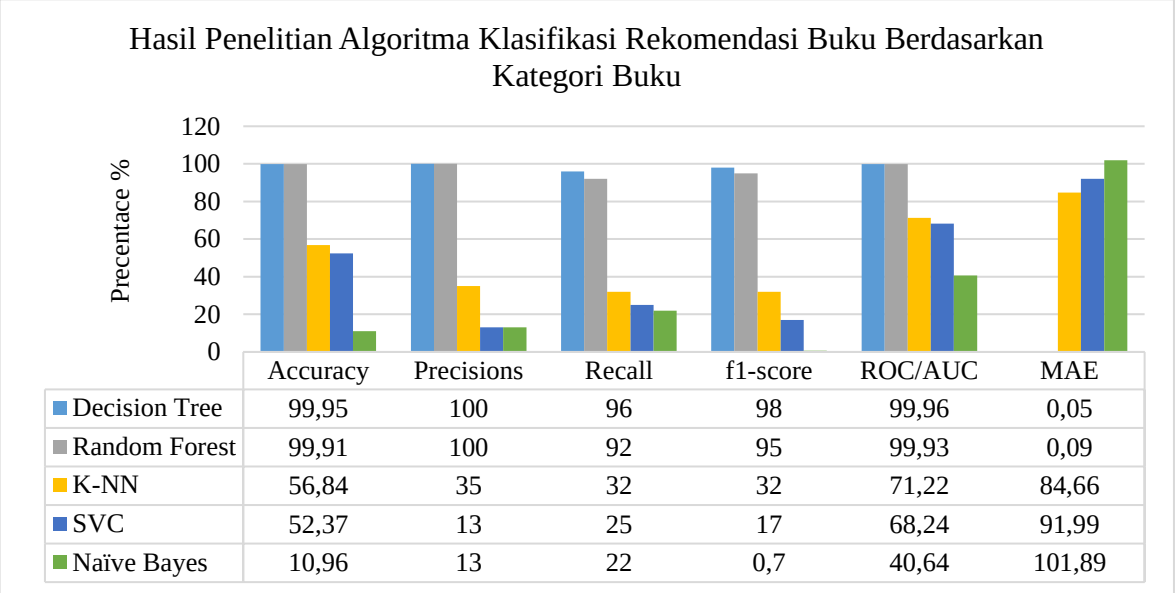
Tabel 2. Hasil Penelitian Perbandingan Algoritma Klasifikasi

Classifier	*Accuracy*	*Precisions*	*Recall*	*f1-score*	*ROC/AUC*	*MAE*
Decision Tree	99.95	100	96	98	99.96	0.05
Random Forest	99.91	100	92	95	99.93	0.09
K-NN	56.84	35	32	32	71.22	84.66
SVC	52.37	13	25	17	68.24	91.99
Naïve Bayes	10.96	13	22	0.7	40.64	101.89

Berdasarkan pada tabel 2, hasil dari penelitian klasifikasi model *Decision Tree* memperoleh nilai akurasi 99.95% dengan nilai *Precisions* 100%, *Recall* 96%, *f1-score* 98%, *ROC/AUC* 99.96% dan rata-rata

p-ISSN : 1979-0155 e-ISSN : 2614-8870

kesalahan sebesar 0.05, hasil tertinggi kedua dengan selisih akurasi yang sedikit 0.04% yaitu model *Random Forest* dengan nilai akurasi 99.91% dan nilai *Precisions* 100%, *Recall* 92%, *f1-score* 95%, *ROC/AUC* 99.93% dan rata-rata kesalahan sebesar 0.09, yang ketiga yaitu dengan selisih yang cukup besar sekitar 43.11% dari nilai akurasi tertinggi yaitu model K-NN dengan nilai akurasi 56.84% disusul dengan model SVC dengan nilai akurasi 52.37% dengan selisih nilai akurasi sebesar 47.58% dari nilai akurasi tertinggi dan yang terakhir yaitu model *Naïve Bayes* sebagai algoritma yang memiliki nilai akurasi terendah dengan nilai akurasi sebesar 10.96%. Dapat uraian hasil diatas dapat disimpulkan bahwa Algoritma *Decision Tree* sebagai algoritma klasifikasi dengan nilai akurasi tertinggi dan menjadi algoritma klasifikasi yang baik untuk dataset *goodreads books*.



	Accuracy	Precisions	Recall	f1-score	ROC/AUC	MAE
Decision Tree	99,95	100	96	98	99,96	0,05
Random Forest	99,91	100	92	95	99,93	0,09
K-NN	56,84	35	32	32	71,22	84,66
SVC	52,37	13	25	17	68,24	91,99
Naïve Bayes	10,96	13	22	0,7	40,64	101,89

Gambar 1. Grafik Hasil Penelitian Perbandingan Algoritma Klasifikasi

Berdasarkan pada grafik Gambar 1. dapat disimpulkan bahwa algoritma klasifikasi yang memiliki *accurancy, precision, recall, f1-score, ROC/AUC* tertinggi yaitu Algoritma *Decision Tree* dengan *Mean Absolute Error* terkecil yaitu 0.05 semakin tinggi nilai akurasi dan AUC, maka semakin tinggi pula untuk memberikan rekomendasi buku berdasarkan kategori buku pada pengguna.

Model *confusion matrix* akan membentuk matrix yang terdiri dari *true positif* atau tupel positif dan *true negatif* atau tupel negatif[10]. Berikut dibawah ini merupakan hasil dari *confusion matrix* dari algoritma klasifikasi *Decision Tree, Random Forest, KNN, SVC* dan *Naïve Bayes*:

Tabel 3. *Confusion Matrix* Algoritma *Decision Tree*

Confusion Matrix Decision Tree			***Predicted***			
			(Best Books)	***(Good Books)***	***(Not-Bad Books)***	***(Lowest Rated Books)***
			0	**1**	**2**	**3**
Actual	***(Best Books)***	**0**	65	0	0	0
	(Good Books)	**1**	0	1137	0	0
	(Not-Bad Books)	**2**	0	0	5	1
	(Lowest Rated Books)	**3**	0	0	0	963

Berdasarkan pada tabel 3, maka yang diprediksi sebagai *Best Books* sebanyak 65 data sesuai dengan prediksi yaitu *Best Books*. Yang diprediksi sebagai *Good Books* memiliki 1137 data sesuai dengan prediksi yaitu *Good Books*. Dan yang diprediksi sebagai *Not-Bad Books* memiliki 5 data yang sesuai dengan prediksi. Sedangkan yang diprediksi sebagai *Lowest Rated Books* memiliki 963 data yang sesuai prediksi, 1 data ternyata *Not-Bad Books*.

Tabel 4. *Confusion Matrix* Algoritma *Random Forest*

			Predicted

Confusion Matrix Random Forest			*(Best Books)*	*(Good Books)*	*(Not-Bad Books)*	*(Lowest Rated Books)*
			0	**1**	**2**	**3**
Actual	*(Best Books)*	**0**	65	0	0	0
	(Good Books)	**1**	0	1137	0	0
	(Not-Bad Books)	**2**	0	0	4	2
	(Lowest Rated Books)	**3**	0	0	0	963

Berdasarkan pada tabel 4, maka yang diprediksi sebagai *Best Books* sebanyak 65 data sesuai dengan prediksi yaitu *Best Books*. Yang diprediksi sebagai *Good Books* memiliki 1137 data sesuai dengan prediksi yaitu *Good Books*. Dan yang diprediksi sebagai *Not-Bad Books* memiliki 4 data yang sesuai dengan prediksi. Sedangkan yang diprediksi sebagai *Lowest Rated Books* memiliki 963 data yang sesuai prediksi, 2 data ternyata *Not-Bad Books*.

Tabel 5. *Confusion Matrix* Algoritma *KNN*

Confusion Matrix KNN			*Predicted*			
			(Best Books)	*(Good Books)*	*(Not-Bad Books)*	*(Lowest Rated Books)*
			0	**1**	**2**	**3**
Actual	*(Best Books)*	**0**	7	38	0	20
	(Good Books)	**1**	17	721	0	399
	(Not-Bad Books)	**2**	0	2	0	4
	(Lowest Rated Books)	**3**	5	452	0	506

Berdasarkan pada tabel 5, maka yang diprediksi sebagai *Best Books* sebanyak 7 data sesuai dengan prediksi yaitu *Best Books*, 17 data ternyata Goods *Books*, 5 data ternyata *Lowest Rated Books*. Yang diprediksi sebagai *Good Books* memiliki 721 data sesuai dengan prediksi yaitu *Good Books*, 38 ternyata *Best Books*, 2 data ternyata *Not-Bad Books*, 452 data ternyata *Lowest Rated Books*. Sedangkan yang diprediksi sebagai *Lowest Rated Books* memiliki 506 data yang sesuai prediksi, 20 data ternyata *Best Books*, 399 data yang ternyata *Good Books* dan 4 data ternyata *Not-Bad Books*.

Tabel 6. *Confusion Matrix* Algoritma *SVC*

Confusion Matrix SVC			*Predicted*			
			(Best Books)	*(Good Books)*	*(Not-Bad Books)*	*(Lowest Rated Books)*
			0	**1**	**2**	**3**
Actual	*(Best Books)*	**0**	0	65	0	0
	(Good Books)	**1**	0	1137	0	0
	(Not-Bad Books)	**2**	0	6	0	0
	(Lowest Rated Books)	**3**	0	963	0	0

Berdasarkan pada tabel 6, maka yang diprediksi sebagai *Good Books* memiliki 65 data ternyata *Best Books*, 1137 sesuai prediksi yaitu *Good Books*, 6 data ternyata *Not-Bad Books*, 963 data ternyata *Lowest Rated Books*.

Tabel 7. *Confusion Matrix* Algoritma *Naïve Bayes*

Confusion Matrix Naïve Bayes			*Predicted*			
			(Best Books)	*(Good Books)*	*(Not-Bad Books)*	*(Lowest Rated Books)*
			0	**1**	**2**	**3**
Actual	*(Best Books)*	**0**	0	6	59	0
	(Good Books)	**1**	2	234	901	0
	(Not-Bad Books)	**2**	0	2	4	0
	(Lowest Rated Books)	**3**	0	220	743	0

Berdasarkan pada tabel 7, maka yang diprediksi sebagai *Best Books* sebanyak 2 data ternyata Goods *Books*. Yang diprediksi sebagai *Good Books* memiliki 6 data ternyata *Best Books*, 234 sesuai prediksi *Good*

p-ISSN : 1979-0155 e-ISSN : 2614-8870

Books, 2 data ternyata *Not-Bad Books*, 220 data ternyata *Lowest Rated Books*. Sedangkan yang diprediksi sebagai *Not-Bad Books* memiliki 4 data yang sesuai prediksi, 59 data ternyata *Best Books*, 901 data ternyata *Good Books*, 743 data ternyata *Lowest Rated Books*.

4. KESIMPULAN DAN SARAN

Dalam penelitian ini melakukan komparasi dari beberapa model algoritma klasifikasi sebagai rekomendasi buku berdasarkan prediksi dari kategori buku pada *Goodreads*. Algoritma dasar yang digunakan yaitu Algoritma *Decision Tre,* sedangkan algoritma yang digunakan sebagai perbandingan yaitu algoritma *Naïve Bayes*, SVM, KNN dan *Random Forest*. Hasil dari evaluasi dan validasi, diketahui bahwa *Decision Tree* memiliki akurasi yang paling tinggi diantara metode yang dikomparasikan sebesar 99,95%, *precision* sebesar 100%, *recall* sebesar 96%, f1-score sebesar 98% dengan rata-rata kesalahan sebesar 0.05 dan AUC sebesar 99,96%, diikuti oleh algoritma *RRandom Forest, K-Nearest Neighbor (K-NN), Support Vector Classifier (SVC).* dan *Naïve Bayes* yan yang memiliki akurasi yang paling rendah. Dengan demikian hasil evaluasi menggunakan curva ROC/AUC yaitu, algoritma klasifikasi *Decision Tree* bernilai 99,96% dengan tingkat diagnose *excellent classification*. Hal itu yang menjadi bukti bahwa algoritma *Decision Tree* dapat digunakan sebagai rekomendasi buku untuk memprediksi kategori buku pada *Goodreads Books*.

Berdasarkan hasil penelitian, saran untuk pengembangan dari penelitian ini adalah Menggunakan metode klasifikasi data mining yang lain seperti metode klasifikasi ensemble (Bagging, AdaBoost) yaitu menggabungkan beberapa metode sebagai solusi klasifikasi untuk mendapatkan hasil terbaik. Melakukan pengembangan dengan *feature selection* seperti *Particle Swarm Optimization, Genetic Algoritmh* dan metode *feature selection* lainnya untuk dapat mengoptimalkan parameter atau menyeleksi atribut yang berpengaruh kuat. Melakukan pengembangan dengan metode deep learning, seperti DNN, ANN, RNN.

DAFTAR PUSTAKA

[1] Goodreads, "GoodReads.com," *[Online]*. https://www.goodreads.com/ (accessed Oct. 20, 2020).

[2] M. Thelwall and K. Kousha, "Goodreads: A social network site for book readers," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 4, pp. 972–983, 2017, doi: 10.1002/asi.23733.

[3] M. A. Ghani and A. Subekti, "Email Spam Filtering Dengan Algoritma Random Forest," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. Vol.3, No., no. 2, p. 216~221, 2018.

[4] A. Franseda, "Integrasi Metode Decision Tree dan SMOTE untuk Klasifikasi Data Kecelakaan Lalu Lintas Integration of Decision Tree and SMOTE Methods for Classification of Traffic Accidents Data," vol. 08, no. 3, 2020, doi: 10.26418/justin.v8i3.40982.

[5] A. M. Maghari, I. A. Al-najjar, S. J. Al-laqtah, and S. S. Abu-naser, "Books ' Rating Prediction Using Just Neural Network," vol. 4, no. 10, pp. 17–22, 2020.

[6] R. A. Tyas, M. Anggraini, I. A. Sulasiyah, and Q. Aini, "Implementasi Algoritma Naïve Bayes Dalam Penentuan Rating Buku," *Sistemasi*, vol. 9, no. 3, p. 557, 2020, doi: 10.32520/stmsi.v9i3.915.

[7] A. Suryanto, I. Alfarobi, and T. A. Tutupoly, "Komparasi Algoritma C4.5, Naive Bayes Dan Random Forest Untuk Klasifikasi Data Kelulusan Mahasiswa Jakarta," *Mitra dan Teknol. Pendidik.*, vol. iv nomor 1, pp. 2–14, 2018, [Online]. Available: https://www.publikasiilmiah.com/jurnal-mitra-dan-teknologi-pendidikan-volume-iv-nomer-1-februari-2018/.

[8] Soumik, "goodreads-books," *www.kaggle.com*, 2020. https://www.kaggle.com/jealousleopard/goodreadsbooks (accessed Oct. 15, 2020).

[9] F. Teknik and U. M. Semarang, "Deteksi Penyakit Algoritma ID3 Gagal Ginjal Kronis Menggunakan," vol. 13, no. 1, pp. 8–17, 2020.

[10] T. A. Tutupoly and I. Alfarobi, "Jurnal Mitra Pendidikan (JMP Online)," *J. Mitra Pendidik.*, vol. 3, no. 1, pp. 11-2292–103, 2019.

[11] M. Lestandy, L. Syafa'ah, and A. Faruq, "Classification of potential blood donors using machine learning algorithms approach," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 3, pp. 217–221, 2020, doi: 10.14710/jtsiskom.2020.13619.