IMPLEMENTASI ALGORITMA *NAÏVE BAYES* DALAM PENENTUAN *RATING* BUKU

Muthia Anggraini, Rizki Ayuning Tyas, Ismi Ana Sulasiyah, Qurrotul Aini
Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah Jakarta,
Jl. Ir H. Juanda No.95, Cemp. Putih, Kec. Ciputat, Kota Tangerang Selatan, 15412
Email: muthia.anggraini17@mhs.uinjkt.ac.id , rizki.ayuning17@mhs.uinjkt.ac.id , ismi.ana17@mhs.uinjkt.ac.id , qurrotulaini@uinjkt.ac.id

(Diterima: 29 Juni 2020 , direvisi: 9 Agustus 2020, disetujui: 12 Agustus 2020)

ABSTRACT

Books are one of the most widely used objects in daily life. With the development of the times, there are other alternatives that can be used to read books without having to buy books in stores. One alternative is the website www.goodreads.com where the website provides a variety of books. On the website, we can also give ratings and review s of books that we have read. These review s and ratings can provide a reference for readers. For this reason, an analysis of book rating is required based on data obtained from the www.kaggle.com website. By processing the data obtained will find the best book viewed from several aspects. The purpose of this research is to determine the rating of a book as a reference for readers in choosing the appropriate book. In this study using a classification algorithm naïve bayes data mining. This research was assisted by rapidminer and Python tools as tools to manage data. The results obtained are the results of determining the book rating using the naïve bayes method having an accuracy of 66.98%, precision 74.47% and recall 62.47% and the results of this analysis are obtained from the dataset available on the website www.kaggle.com showing that the majority book rating predictions tend to be low.

Keywords: *book rating, mining, naïve bayes, pyhton, rapidminer*

ABSTRAK

Buku merupakan salah satu benda yang paling banyak digunakan dalam kehidupan sehari-hari. Dengan berkembangnya zaman, ada alternatif lain yang bisa digunakan untuk membaca buku tanpa harus membeli buku ditoko. Salah satu alternatifnya adalah *website* www.goodreads.com yang dimana *website* tersebut menyediakan berbagai macam buku. Di *website* tersebut, kita juga dapt memberikan *rating* dan *review* buku yang telah kita baca. *Review* dan *rating* ini bisa memberikan acuan bagi para pembaca. Untuk itu diperlukan nya analisis terhadap penentuan *rating* buku berdasarkan data yang didapatkan dari situs www.kaggle.com. Dengan mengolah data yang didapatkan akan mengetahui buku yang paling terbaik dilihat dari beberapa segi. Adapun tujuan dilakukan penelitian ini adalah untuk menentukan *rating* dari sebuah buku sebagai acuan pembaca dalam memilih buku yang sesuai. Dalam penelitian ini menggunakan algoritma klasifikasi data *mining naïve bayes.* Penelitian ini dibantu oleh *tools rapidminer* dan *Python* sebagai alat bantu mengelolah data. Hasil yang diperoleh adalah hasil penentuan *rating* buku menggunakan metode *naïve bayes* memiliki *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.

Kata Kunci: *rating* buku , *mining*, *naïve bayes*, *python*, *rapidminer*

1 PENDAHULUAN

Menurut Kamus Besar Bahasa Indonesia, Buku adalah lembar kertas yg berjilid, berisi tulisan atau kosong [1]. Buku adalah kumpulan kertas atau bahan lainnya yang dijilid menjadi satu pada salah satu ujungnya dan berisi tulisan atau gambar [1]. Setiap sisi dari sebuah lembaran kertas pada buku disebut sebuah halaman. Banyak jenis-jenis buku yaitu, majalah, novel, komik, kitab suci, dan juga naskah. Seiring dengan perkembangan dalam bidang dunia informatika, kini dikenal pula istilah *e-book* (buku elektronik), yang mengandalkan perangkat seperti komputer meja, komputer jinjing,

komputer tablet, telepon seluler dan lainnya, serta menggunakan perangkat lunak tertentu untuk membacanya. Buku cetak pada umumnya terdiri atas setumpuk kertas dijilid yang berisi teks atau gambar, maka buku elektronik berisikan informasi digital yang dapat berisi teks, gambar, *audio, video*.

Pada era sekarang, untuk membaca buku tidaklah harus membeli ditoko buku. Sekarang sudah banyak *platform* atau *website* yang menyediakan untuk bisa membaca buku secara berbayar ataupun gratis dan buku tersebut tidak perlu digenggam atau dipegang oleh kita. Cukup dengan membawa *handphone* ataupun perangkat keras lainnya yang bisa terhubung dengan internet, maka kita dapat membaca buku. Salah satunya adalah *website* dari *www.goodreads.com* yang menyediakan bacaan buku novel yang bisa dibaca secara *online* dan gratis. Tapi tidak semua buku bisa dibaca disana, walau begitu cukup banyak buku yang tersedia disana. Selain bisa membaca, disana kita juga bisa memberikan *review* dan *rating* untuk setiap buku yang kita baca. *Review* dan *rating* ini bisa menjadi masukan untuk para penulis buku tersebut.

Data *mining* merupakan sebuah metode yang digunakan dalam pengolahan data berskala besar oleh karena itu data *mining* memiliki peranan yang sangat penting dalam beberapa bidang kehidupan diantaranya yaitu bidang industri, bidang keuangan, cuaca, ilmu dan teknologi. Data *mining* juga bisa diartikan sebagai rangkaian kegiatan untuk menemukan pola yang menarik dari data dalam jumlah besar, kemudian data – data tersebut dapat disimpan dalam *database*, data *warehouse* atau penyimpanan informasi [7].

Salah satu teknik yang dibuat dalam data *mining* adalah adalah bagaimanana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan. Dalam data *mining*, pengelompokan data juga bisa dilakukan. Tujuannya adalah agar kita dapat mengetahui pola universal data-data yang ada. Ada beberapa teknik yang dimiliki data *mining* berdasarkan tugas yang bisa dilakukan, yaitu [8]: Pertama, klasifikasi. Dalam klasifikasi, terdapat target *variabel* kategori. Dalam klasifikasi *variabel*, tujuan bersifat kategorik. Misalnya, kita akan mengklasifikasikan pendapatan dalam tiga kelas, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Kedua, clustering. Clustering dapat dipakai untuk memberikan label pada kelas data yang belum diketahui. Clustering lebih ke arah pengelompokan *record*, pengamatan, atau kasus dalam kelas yang memiliki kemiripan. Ketiga, asosiasi. Asosiasi adalah teknik *mining* untuk menemukan asosiatif antara kombinasi atribut. Mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada satu waktu.

Salah satu metode klasifikasi dari data *mining* adalah *naive* bayes. *Naïve bayes* merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari *dataset* yang diberikan [9]. Algoritma mengunakan *teorema Bayes* dan mengasumsikan semua atribut *Independen* atau tidak saling ketergantungan yang diberikan oleh nilai pada *variabel* kelas [10]. Keuntungan penggunaan *Naïve bayes* adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (*Training* Data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. *Naïve bayes* sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan [11].

Untuk menganalisis *review* dan *rating* yang sudah ada, kita bisa mengimplementasikan salah satu metode dari data *mining*, yaitu *naïve bayes*. Metode *naïve bayes* ini digunakan untuk mengklasifikasikan beberapa informasi sejenis menjadi satu kesatuan. Metode *naïve bayes* adalah metode yang mempunyai tingkat kesalahan yang sangat minimum jika dibandingkan algoritma klasifikasi lainnya [2]. Walaupun memiliki tingkat kesalahan yang sangat minumum, metode *naïve bayes* ini memiliki kekurangan yaitu rendahnya tingkat kepekaan terhadap data yang hilang atau *missing value*. Oleh karena itu tetap diperlukan pengukuran tingkat ke akuratan untuk perhitungan dengan menggunakan data ini dengan cara *confusion matrix*.

Pada penelitian ini akan dilakukan penelitian menggunakan metode *naïve bayes* yang dibantu oleh *software rapidminer*. Yang dimana kami melakukan penelitian ini untuk menentukan prediksi *rating* buku yang terdapat di *www.goodreads.com*. Berdasarkan permasalahan yang telah dipaparkan sebelumnya, adapun tujuan dilakukan penelitian ini adalah untuk menentukan *rating* dari sebuah buku sebagai acuan pembaca dalam memilih buku yang sesuai pada *website* *www.goodreads.com*. Peneliti menggunakan data "*books listed in goodreads"* yang didapat dari situs *www.kaggle.com*. Hasil akhir yang diberikan berupa hasil *rating* menggunakan metode *naïve bayes*.

2 TINJAUAN PUSTAKA

Berdasarkan hasil penelitian sebelumnya [3], penelitian ini membahas tentang bagaimana mengimplementasikan disiplin ilmu data *mining* menggunakan komparasi metode *naïve bayes* dengan algoritma C4.5 yang merupakan sebuah metode untuk melakukan teknik klasifikasi serta diaplikasikan dengan *tools rapidminer*. Penelitian ini memperoleh hasil bahwa dengan pemilihan data *training* 50 *record*, 4 atribut *predictor* dan 1 atribut target menghasilkan 5 aturan dalam pohon keputusan sehingga aturan tersebut dapat digunakan dalam menentukan kelulusan tepat waktu pada mahasiswa STMIK STIKOM BALI. Hasil analisis menggunakan metode *naïve bayes* diperoleh hasil akurasinya sebesar 89,27% dimana hasil *performance* akurasi menunjukan kelulusan tepat waktu sebanyak 40 dan tidak tepat 10.

Pada penelitian selanjutnya yang berkaitan dengan penelitian ini menyatakan bahwa dengan adanya *review* terhadap sebuah film bisa membantu penonton untuk lebih selektif lagi dalam memilih suatu film. Dan dari pihak produksi bisa terbantu untuk mengukur seberapa jauh kualitas film yang mereka hasilkan. Penelitian tersebut menunjukan bahwa pihak produksi sendiri terkadang mengalami kesulitan dalam memilah dan mengkategorikan *review*, apakah produk tersebut kualitasnya tergolong bagus, cukup bagus, tidak bagus, dan lainnya. Dalam penelitian in penilaian suatu film berdasarkan *review* yang telah diberikan *rating*. Penelitian menggunakan metode klasifikasi *naïve bayes* yang dimana hasil klasifikasi yang didapat tersebut kemudian dihitung nilai *accuracy*, *precision*, *recall*. Yang dimana *accuracy* 55,80%, *precision* 32,41% dan *recall* 46,70 %. Pada penelitian ini hasil untuk *naïve bayes* pada model distribusi untjk nilai *class* "RENDAH" sebanyak 0,707, sedangkan *class* "TINGGI" sebanyak 0,293, menunjukan penentuan *rating* film (*imdb_score*) tinggi apabila tahun film (*title_year*) baru serta *duration* film panjang [4].

Mengacu pada penelitian berikutnya [5] yang membahas tentang pengklasifikasian teks otomatis pengaduan dan pelaporan masyarakat yang ada di Kepolisian Negara Republik Indonesia menggunakan metode *naïve bayes classifier*. Yang menjadi probabilitas kata kunci dalam penelitian ini adalah membandingkan dokumen latih dan dokumen uji. Keduanya dibandingkan dengan beberapan tahapan persamaan yang pada akhirnya akan diperoleh hasil tertinggi untuk ditetapkan sebagai kategori dokumen baru. Untuk hasil penelitian ini, pengklasifikasian teks otomatis pelaporan dan pengaduan masyarah menghasilkan rata-rata akurasi yang tinggi, yaitu dengan *recall* 93%, *precission* 90%, dan *f-measure* 92%.

Penelitian lainnya [6] membahas bagaimana algoritma *naïve bayes* dapat membantu untuk memprediksikan angka kelahiran. Penelitian ini dilakukan didaerah Medan tepatnya di Kabupaten Batubara yang dimana terdiri dari 10 dusun. Tujuan penelitian ini adalah melihat pola prediksi dari setiap atribut-atribut yang terdapat pada *dataset* dengan menggunakan algoritma *naïve bayes*. Penelitian ini dibantu dengan bahasa pemrograman *visual basic 2008* serta *microsoft access 2007* sebagai basis data. Hasil dari penelitian ini adalah mempermudah pihak kantor desa lalang dalam proses pengelolaan data penduduk, membantu dalam proses penginputan data, pencarian data dan laporan penduduk.

Pada penelitian ini, akan dilakukan penentuan *rating* dari sebuah buku yang akan dijadikan acuan pembaca dalam memilih buku yang sesuai. Berdasarkan beberapa penelitian sebelumnya, implementasi algoritma *naïve bayes* digunakan untuk mengklasifikasikan data menjadi beberapa kelas. Peneliti menggunakan metode *naïve bayes* karena metode ini membantu penelitian ini dalam mengklasifikasikan *rating* dari sebuah buku. Penelitian ini dibantu dengan menggunakan *software rapidminer* dan *Python* versi 3.7 yang mempermudah dalam mengklasifikasikan *dataset* yang digunakan.

3 METODE PENELITIAN

Berikut adalah uraian metode yang digunakan selama penelitian. Selama penelitian ini berlangsung terdapat 2 metode yang digunakan, yaitu metode untuk pengumpulan data dan metode untuk pengelolaan data.

Pengumpulan data dalam penelitian ini diambil berdasarkan data-data sampel dari situs internet, yaitu situs *www.kaggle.com*. Data yang didapatkan adalah data berupa data statistik buku-buku pada *website* *www.goodreads.com*. Data yang terkumpul sebanyak 11.127 yang terdiri dari atribut *bookID, title. author, average_rating, isbn, isbn13, language_code, num_pages, rating_count,*

text_review _count, publication_date, publisher yang kemudian akan diolah kembali sesuai dengan batasan atau ketentuan penelitian.

Pada penelitian ini, pengolahan data dilakukan dengan menggunakan bantuan *tools* berupa *software rapidminer*, yang dimana *rapidminer* adalah salah satu *software* untuk melakukan pengolahan data *mining*. Selain *rapidminer*, pengolahan data juga dibantu dengan bahasa pemrograman Pyhton. Langkah-langkah yang dilakukan dalam pengolahan data adalah sebagai berikut:

1. *Selection*
   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*. Data *training* adalah data yang digunakan untuk perhitungan probabilitas dari data berdasarkan data pembelajaran yang dilakukan. Sedangkan data *testing* merupakan data yang akan atau sedang terjadi dan dipergunakan sebagai bahan uji yang sebelumnya sudah didapatkan pada data *training*.
2. *Pre-Processing*
   Tahapan *pre-processing* dilakukan karena ketika data dimasukan kedalam *software rapidminer* terdapat *missing value* yang berarti data yang didapat belum normal. Kemudian didalam tahapan *pre-processing* ini dilakukan tahap untuk mengurangin atribut-atribut yang masih terdapat *missing value* dengan tahapan transformasi yang dilakukan dengan *software rapidminer.*
3. *Data Mining*
   Pada penelitian ini, proses data *mining* dilakukan menggunakan metode *naïve bayes* untuk melakukan proses analisis pada *dataset* "*books listed in goodreads*" yang didapat di situs *www.kaggle.com*. Tahapan ini adalah tahapan dimana dilakukannya proses *training* pada data-data yang akan digunakan, dengan tujuan untuk membuat melatih model algoritma yang akan digunakan untuk data *testing*.
4. Pengujian Model dan *Evaluation*
   Tahapan ini adalah tahapan dimana data yang sudah melalui proses pelatihan (*training*) akan ditentukan akurasi kualitas datanya sekaligus melakukan *testing* terhadap model yang telah dibuat. Dengan menggunakan metode *confusion matrix*, data yang dimana tidak digunakan untuk *training* akan digunakan menjadi data test untuk dilakukannya *testing* terhadap model algoritma yang sudah dibuat sekaligus menentukan akurasi dari hasil analisis. Yang dimana untuk pengujian *Accuracy*, *Precission*, *Re-call* agar bisa membuktikan kinerja metode *naïve bayes* menggunakan persamaan sebagai berikut [12]:
   a. *Accuracy*

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)}\ x100\% \quad (1)$$

   b. *Precission*

$$Precission = \frac{TP}{TP+FP}\ x\ 100\% \quad (2)$$

   c. *Re-call*

$$Recall\ = \frac{TP}{TP+FN}\ x\ 100\% \quad (3)$$

4 HASIL DAN PEMBAHASAN

Berikut adalah hasil pengolahan data yang telah dikerjakan, sebagai berikut:

4.1 Hasil *Selection*

Dataset yang digunakan adalah *dataset* "*books listed in goodreads"* yang diambil dari *website* *www.kaggle.com*. Selama tahun 1900-2019, dalam *dataset* ini diberikan jumlah sebanyak 11.127 buku. Adapun atribut yang digunakan pada penelitian ini terbagi menjadi dua jenis *variabel*, antara lain;

1. *Variabel Dependen* (Y)
   Variabel ini disebut juga dengan nilai yang terikat. *Variabel* Y dalam penelitian ini adalah *average_rating* dimana *variabel* ini merupakan nilai untuk memprediksi *rating* buku.
2. *Variabel Independen* (X)
   Variabel ini merupakan *variabel* dengan nilai yang tidak bergantung dengan *variabel* lainnya. *Variabel* ini bisa disebut juga *variabel* yang kehadirannya dibutuhkan oleh *variabel* terikat. Nilai *class* untuk *variabel* X dapat dilihat pada Tabel 1.

Tabel 1. Nilai *Class* Untuk Setiap Atribut

No.	Nama Atribut	Class
1	*bookID*	*Integer*
2	*Title*	*Text*
3	*Author*	*Text*
4	*Isbn*	*Integer*
5	*isbn13*	*Integer*
6	*language_code*	*Text*
7	*num_pages*	Tinggi ≥ 1000, Sedang ≥ 500, Rendah < 500
8	*rating_counts*	Tinggi ≥ 100000, Sedang ≥ 10000, Rendah < 10000
9	*text_review _count*	Tinggi ≥ 5000, Sedang ≥ 500, Rendah < 500
10	*publication_date*	*Text*
11	*Publisher*	*Text*

Dalam tahap ini juga, *dataset* akan dibagi menjadi data *training* dan data *testing*. *Training-Set* ini akan digunakan untuk membuat model *machine learning*, sedangkan *Test-Set* akan digunakan untuk menguji performa dan kebenaran (terhadap korelasi) dalam model yang bersangkutan. Metode yang digunakan untuk pembagian ini menggunakan peneliti menggunakan *Python* versi 3.7 dengan membagi data yang telah dimuat menjadi dua berdasarkan ukuran standar pembagian *dataset* yaitu, 80% diantaranya akan digunakan untuk melatih model dan 20% digunakan untuk data validasi. Tabel 2 dan Tabel 3 masing-masing menunjukan potongan hasil dari pembagian kedua data ini.

E-ISSN:2540-9719
ISSN:2302-8149

Tabel 2. Data *Training* Dengan Jumlah Data 8901

average_rat...	bookID	title	authors	isbn	isbn13	language_c...	num_pages	ratings_count	text_rev
4.13	14492	Farewell to S...	Leon Sciaky/...	1589680021	9.78159E+12	eng	299	63	3
3.88	4135	I Like You: Ho...	Amy Sedaris	446578843	9.78045E+12	eng	304	37258	1401
4.6	37976	The Listener'...	Anonymous/...	1931047170	9.78193E+12	eng	77	28	3
4.14	17255	McCoy: The P...	David R. Geo...	743491688	9.78074E+12	eng	640	332	38
3.84	31486	Ruby the Red...	Daisy Meado...	043973861X	9.78044E+12	en-US	65	4334	319
3.89	6691	My Uncle Os...	Roald Dahl	140055770	9.78014E+12	eng	208	9170	620
3.41	22944	Drum Into Sil...	Jo Clayton/Ke...	812551249	9.78081E+12	en-GB	399	1	0
3.95	7665	Travels	Michael Crich...	60509958	9.78006E+12	eng	400	7084	514
3.44	22348	The Mystery P...	Grant Morriso...	1563891891	9.78156E+12	en-US	80	688	40
3.98	764	Del amor y ot...	Gabriel Garcí...	307350444	9.78031E+12	spa	176	4508	278
3.74	17181	CliffsNotes o...	James Lama...	822012197	9.78082E+12	en-GB	72	30	9
3.47	36333	Loving Will S...	Carolyn Meyer	152054510	9.78015E+12	eng	265	1361	166
4.35	35729	Lover Eternal ...	J.R. Ward	451218043	9.78045E+12	eng	464	155348	5325

ExampleSet (8.901 examples, 1 special attribute, 11 regular attributes)

Tabel 3. Data *Testing* Dengan Jumlah Data 2226

average_rat...	bookID	title	authors	isbn	isbn13	language_c...	num_pages	ratings_count	text_rev
3.69	33896	Thirteen Moo...	Charles Frazier	375509321	9.78038E+12	eng	422	10534	1578
3.43	6480	Waverley	Walter Scott/...	192836013	9.78019E+12	eng	463	69	8
4.43	44350	Beautiful Boy...	Francesca Li...	60594357	9.78006E+12	en-US	304	413	5
3.9	15910	Geisha: The ...	Jodi Cobbla...	037570180X	9.78038E+12	eng	128	188	12
3.8	37326	Fragments	Jean Baudrill...	1844675734	9.78184E+12	eng	148	127	11
3.99	24048	The Abductio...	Gordon Korm...	043984777X	9.78044E+12	eng	144	3306	263
4.22	13914	The State of ...	Sayadaw U. ...	861713451	9.78086E+12	eng	170	28	1
4.44	36681	Drop The Ro...	Bill Pittman/T...	1592851614	9.78159E+12	en-US	132	955	41
3.6	7113	The King in th...	Adam Gopnik...	786838949	9.78079E+12	en-US	410	805	108
3.87	16180	The Boleyn In...	Philippa Greg...	743272501	9.78074E+12	eng	518	68763	3394
3.52	31178	Back When ...	Anne Tyler	345477243	9.78035E+12	eng	336	15040	1239
3.88	9593	Galápagos	Kurt Vonnegu...	385333870	9.78039E+12	eng	324	55850	1990
4.14	40424	Tarzan of the ...	Edgar Rice B...	517659573	9.78052E+12	eng	848	29	1

ExampleSet (2.226 examples, 1 special attribute, 11 regular attributes)

4.2 Hasil *Pre-processing*

Dalam tahap *pre-processing* ini akan dilakukan identifikasi terhadap data yang tidak konsisten, duplikasi data, dan data yang tidak lengkap (*missing value*). Proses ini dilakukan dengan seleksi data-data yang tidak konsisten dan penghapusan *missing value*. *Dataset* yang dimiliki peneliti tidak memiliki data yang tidak konsisten secara signifikan dan hanya mengurangi sedikit data yang dimiliki menjadi 11.124 baris. Jumlah data yang dimiliki setelah tahap *pre-processing* ini selanjutnya akan digunakan untuk pembagian data dalam bentuk data *training* dan data *testing*.

Setelah membagi data menjadi data *training* dan data *testing*, akan dilakukan metode *cleaning* supaya data yang digunakan valid sesuai kebutuhan. Sehingga dari nilai *class* data buku dalam atribut tidak terjadi ketidakkonsistenan data dalam pengujian. Lalu data tersebut melalui pengolahan, dalam pengolahan data peneliti menggunakan *software rapidminer studio* versi 9.6 yang nantinya dapat menghasilkan sebuah prediksi *rating* buku. Tabel 4 menunjukkan potongan data *training* sebelum tahap *pre-processing*, sedangkan Tabel 5 merupakan potongan data *training* sesudah tahap *pre-processing*.

E-ISSN:2540-9719
ISSN:2302-8149

Tabel 4. Data *Training* Sebelum *Pre-Processing*

average_rat...	bookID	title	authors	isbn	isbn13	language_c...	num_pages	ratings_count	text_rev
4.13	14492	Farewell to S...	Leon Scaky/...	1589880021	9.78159E+12	eng	299	63	3
3.88	4135	I Like You: Ho...	Amy Sedaris	446578843	9.78045E+12	eng	304	37258	1401
4.6	37976	The Listener'...	Anonymous/...	1931047170	9.78193E+12	eng	77	28	3
4.14	17255	McCoy: The P...	David R. Geo...	743491688	9.78074E+12	eng	640	332	38
3.84	31486	Ruby the Red...	Daisy Meado...	043973861X	9.78044E+12	en-US	65	4334	319
3.89	6691	My Uncle Os...	Roald Dahl	140055770	9.78014E+12	eng	208	8170	629
3.41	22944	Drum Into Sil...	Jo Clayton/Ke...	812551249	9.78081E+12	en-GB	399	1	0
3.95	7665	Travels	Michael Crich...	60509058	9.78006E+12	eng	400	7084	514
3.44	22348	The Mystery P...	Grant Morriso...	1563891891	9.78156E+12	en-US	80	688	40
3.98	764	Del amor y ot...	Gabriel Garcí...	307350444	9.78031E+12	spa	176	4508	278
3.74	17181	CliffsNotes o...	James Lama...	822012197	9.78082E+12	en-GB	72	30	9
3.47	36333	Loving Will S...	Carolyn Meyer	152054510	9.78015E+12	eng	265	1361	166
4.35	35729	Lover Eternal ...	J.R. Ward	451218043	9.78045E+12	eng	464	155348	5325

ExampleSet (8.901 examples, 1 special attribute, 11 regular attributes)

Tabel 5. Data *Training* Sesudah *Pre-Processing*

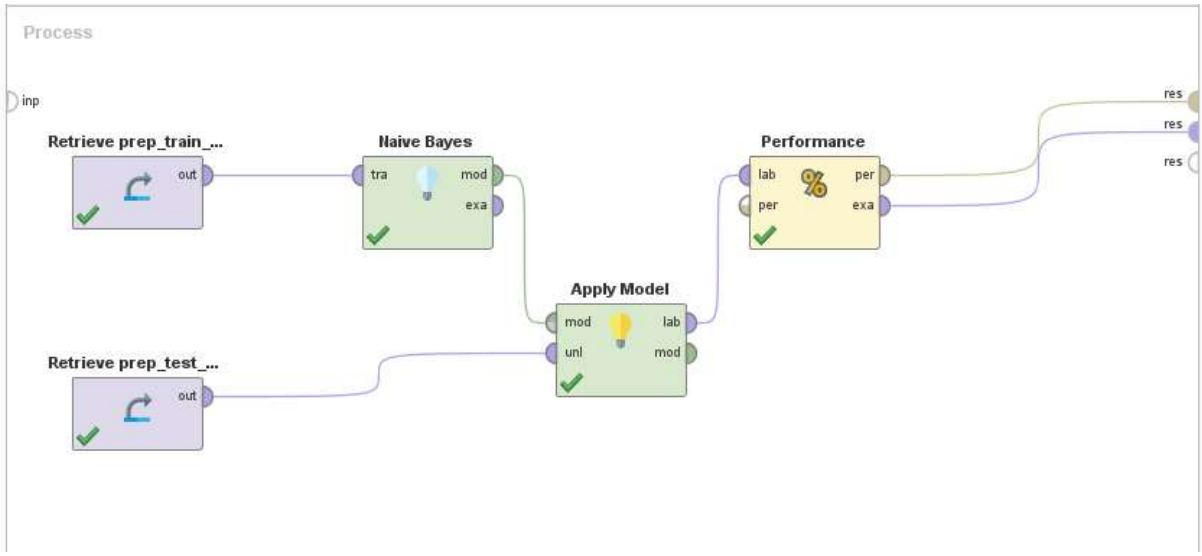
Open in | Turbo Prep | Auto Model | Filter (8.898 / 8.898 examples): all

average_rat...	title	authors	isbn	isbn13	language_c...	publication_...	publisher	bookID	num_pa
Tinggi	Farewell to S...	Leon Scaky/...	1589880021	9.78159E+12	eng	5/1/2003	Paul Dry Books	14492	Rendah
Rendah	I Like You: Ho...	Amy Sedaris	446578843	9.78045E+12	eng	10/16/2006	Warner Book...	4135	Rendah
Tinggi	The Listener'...	Anonymous/...	1931047170	9.78193E+12	eng	1/1/2001	Fellowship fo...	37976	Rendah
Tinggi	McCoy: The P...	David R. Geo...	743491688	9.78074E+12	eng	8/29/2006	Pocket Books	17255	Sedang
Rendah	Ruby the Red...	Daisy Meado...	043973861X	9.78044E+12	en-US	5/1/2005	Scholastic Inc.	31486	Rendah
Rendah	My Uncle Os...	Roald Dahl	140055770	9.78014E+12	eng	5/1/1986	Penguin (No...	6691	Rendah
Rendah	Drum Into Sil...	Jo Clayton/Ke...	812551249	9.78081E+12	en-GB	2/1/2004	Tor Fantasy	22944	Rendah
Rendah	Travels	Michael Crich...	60509058	9.78006E+12	eng	11/5/2002	Harpperen	7665	Rendah
Rendah	The Mystery P...	Grant Morriso...	1563891891	9.78156E+12	en-US	9/1/1995	Vertigo	22348	Rendah
Rendah	Del amor y ot...	Gabriel Garcí...	307350444	9.78031E+12	spa	2/7/2006	Plaza y Janes	764	Rendah
Rendah	CliffsNotes o...	James Lama...	822012197	9.78082E+12	en-GB	10/3/1963	Cliffs Notes	17181	Rendah
Rendah	Loving Will S...	Carolyn Meyer	152054510	9.78015E+12	eng	10/1/2006	Harcourt Chil...	36333	Rendah
Tinggi	Lover Eternal ...	J.R. Ward	451218043	9.78045E+12	eng	3/7/2006	Signet	35729	Rendah

ExampleSet (8.898 examples, 1 special attribute, 11 regular attributes)

4.3 Hasil Data *Mining*

Pada penelitian ini, proses data *mining* dilakukan menggunakan metode *naïve bayes* untuk melakukan proses analisis pada *dataset* buku. Gambar 1 merupakan rancangan proses klasifikasi *naïve bayes* yang dilakukan menggunakan *software rapidminer.*



Gambar 1. Klasifikasi *Naïve Bayes*

4.4 Hasil Pengujian Model dan *Evaluation*

Rapidminer sebagai solusi untuk memprediksi dan menganalisis komputasi statistik [13]. Model yang telah dibentuk diuji tingkat akurasinya dengan memasukkan data uji yang berasal dari data *training*. Pada Gambar 2 menunjukan distribusi *class* untuk label atribut *average_rating*. Hasil pada Gambar 2 menunjukan model distribusi *naïve bayes*. Pada hasil *naïve bayes* bisa dilihat bahwa model distribusi nilai *class* "Tinggi" sebanyak 0,447, sedangkan *class* "Rendah" sebanyak 0,553.

SimpleDistribution

```
Distribution model for label attribute average_rating


Class Tinggi (0.447)
11 distributions

Class Rendah (0.553)
11 distributions
```

Gambar 2. Model Distribusi *Naïve Bayes*

Setelah melakukan perhitungan untuk metode *naïve bayes*, maka kita melakukan metode pengujian untuk akurasi data hasil perhitungan tadi dengan metode *confusion matrix. Confusion matrix* adalah suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep *data mining* [14]. *Confusion matrix* memberikan penilaian *performance* klasifikasi berdasarkan objek dengan benar atau salah [7].

E-ISSN:2540-9719
ISSN:2302-8149

Tabel 6. Tingkat *Accuracy, Precision, Recall* Pada *Naïve Bayes*

accuracy: 66.98%

	true Rendah	true Tinggi	class precision
pred. Rendah	779	267	74.47%
pred. Tinggi	468	712	60.34%
class recall	62.47%	72.73%	

Berdasarkan Tabel 6 nilai didapat bahwa *accuracy* 66.98%, *precision* 74.47% dan *recall* 62.47%. Selain menggunakan *rapidminer*, untuk menguji nilai tersebut kita juga dapat menghitung nilai *accuracy*, *precision*, dan *recall* secara manual, yaitu sebagai berikut:

1. *Accuracy*

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}\ x\ 100\%$$
$$Accuracy = \frac{779 + 712}{779 + 267 + 712 + 468}\ x100\%$$
$$Accuracy = \ 0.6698\ x\ 100\%$$
$$Accuracy = \ 66.98\%$$

2. *Precission*

$$Precission\ = \frac{TP}{TP + FP}\ x\ 100\%$$
$$Precission = \frac{779}{779 + 267}\ x\ 100\%$$
$$Precission = 0.7447\ x\ 100\%$$
$$Precission = 74.47\%$$

3. *Recall*

$$Recall\ = \frac{TP}{TP + FN}\ \text{x}\ 100\%$$
$$Recall = \frac{779}{779 + 468}\ x\ 100\%$$
$$Recall = \ 0.6247\ x\ 100\%$$
$$Recall = 62.47\%$$

5 KESIMPULAN

Model *naïve bayes* memiliki tingkat kesalahan yang sangat minimum jika dibandingkan dengan algoritma klasifikasi lainnya. Penelitian yang dilakukan adalah penentuan *rating* buku. Pada hasil *naïve bayes* bisa dilihat bahwa model distribusi nilai *class* "Tinggi" sebanyak 0,447, sedangkan *class* "Rendah" sebanyak 0,553. Hasil penelitian menentukan bahwa hasil penentuan *rating* buku menggunakan metode *naïve bayes* memiliki *accuracy* 66.98%, *precision* 74.47% dan *recall* 62.47%. Berdasarkan hasil penelitian yang didapatkan menggunakan *dataset* dari situs *https://kaggle.com/* menunjukan bahwa *rating* buku pada data *testing* sebanyak 2226 data mayoritas prediksi *rating* buku cenderung rendah. Untuk penelitian selanjutnya, disarankan untuk menggunakan variable yang lainnya yang bisa menjadi bahan pertimbangan dan menggunakan metode data *mining*

lainnya seperti *Decision Trees*, atau *Neural Network* agar bisa menjadi pembanding untuk metode *naïve bayes*.

REFERENSI

[1] H. Alwi, *Kamus Besar Bahasa Indonesia*, 3rd ed. Jakarta: Balai Pustaka, 2007.
[2] J. Liu, Z. Tian, P. Liu, J. Jiang, and Z. Li, "An Approach of Semantic Web Service Classification Based on Naive Bayes," in *2016 IEEE International Conference on Services Computing (SCC)*, San Francisco, CA, USA, Jun. 2016, pp. 356–362, doi: 10.1109/SCC.2016.53.
[3] N. L. Ratniasih, "Optimasi Data *Mining* Menggunakan Algoritma *Naïve bayes* Dan C4.5 Untuk Klasifikasi Kelulusan Mahasiswa," *J. Teknol. Inf. Dan Komput.*, vol. 5, no. 1, Feb. 2019, doi: 10.36002/jutik.v5i1.634.
[4] T. M. Butar, M. A. Fauzi, and Indriati, "Penentuan *Rating Review* Film Menggunakan Metode Multinomial *Naïve bayes* Classifier dengan Feature Selection berbasis Chi-Square dan Galavotti-Sebastiani-Simi Coefficient," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, vol. 3, pp. 447–453, 2019.
[5] F. Handayani and F. S. Pribadi, "Implementasi Algoritma *Naïve bayes* Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110," *J. Tek. Elektro*, vol. 7, pp. 19–24, Jun. 2015.
[6] M. Idris, "Implementasi Data *Mining* Dengan Algoritma *Naïve bayes* Untuk Memprediksi Angka Kelahiran," *J. Pelita Inform.*, vol. 18, pp. 160–167, 2019.
[7] F. Gorunescu, *Data Mining Concepts Models and Techniques*, 1st ed., vol. 12. Craiova: Springer Berlin Heidelberg, 2011.
[8] M. Ridwan and H. Suyono, "Penerapan Data *Mining* Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma *Naïve bayes* Classifier," vol. 7, no. 1, p. 6, 2013.
[9] A. Saleh, "Implementasi Metode Klasifikasi *Naïve bayes* Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," vol. 2, no. 3, p. 11, 2015.
[10] T. R. Patil and M. S. S. Sherekar, "Performance Analysis of *Naïve bayes* and J48 Classification Algorithm for Data Classification," *Open Access*, vol. 6, p. 6, 2013.
[11] A. Powar and D. V. Ghorpade, "Heart Disease Prediction System Using *Naïve bayes* Data *Mining* Technique," *ICTACT J. SOFT Comput.*, pp. 1824–1830, 2018.
[12] T. B. Sasongko, "Komparasi dan Analisis Kinerja Model Algoritma SVM dan PSO-SVM (Studi Kasus Klasifikasi Jalur Minat SMA)," *J. Tek. Inform. Dan Sist. Inf.*, vol. 2, no. 2, Aug. 2016, doi: 10.28932/jutisi.v2i2.476.
[13] M. Utmal and R. K. Pandey, "Taxonomy on the Integration of Hadoop and Rapid Miner for Big Data Analytics," in *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, Jabalpur, India, Dec. 2015, pp. 890–893, doi: 10.1109/CICN.2015.175.
[14] M. F. Rahman, D. Alamsah, M. I. Darmawidjadja, and I. Nurma, "Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN)," *J. Inform.*, vol. 11, no. 1, p. 36, Jan. 2017, doi: 10.26555/jifo.v11i1.a5452.