Prediksi Rating Film Menggunakan Metode Naïve Bayes

Riszki Wijayatun Pratiwi[1], Yusuf Sulistyo Nugroho[2]
Program Studi Informatika, Fakultas Komunikasi Dan Informatika, Universitas Muhammadiyah Surakarta
[1]riszkiwp@gmail.com, [2]*yusuf.nugroho@ums.ac.id*

ABSTRAK

Pada saat ini perkembangan dunia perfilman sudah sangat pesat, contohnya dengan banyaknya film-film yang silih berganti untuk ditayangkan. Namun film-film yang ada tidak semuanya dapat dinikmati dan tidak semua kalangan menyukai semua film. Agar suatu film dapat terus berkembang, tentunya membutuhkan penilaian-penilaian dari para penikmat film, untuk mengetahui selera film yang sesuai dengan para penikmat film. Untuk itu dibutuhkan analisis agar dapat mengetahui bagaimana minat penikmat film yaitu dengan membuat penilaian-penilaian yang nantinya digunakan untuk mengetahui rating suatu film menggunakan metode naïve bayes yaitu metode yang melakukan pendekatan matematika yang fundamental dalam pengenalan pola (pattern recognition). Pendekatan ini didasarkan pada kuantifikasi trade-off antara baerbagai keputusan klasifikasi dengan menggunakan probabilitas dan resiko yang ditimbulkan dalam keputusan-keputusan tersebut. Metode tersebut merupakan salah satu metode dari data mining, dengan atribut yang sudah ditentukan, yaitu meliputi genre film, aktor film, bahasa,warna, durasi film, negara, dan lainnya yang dapat digunakan sebagai tolak ukur sutradara untuk membuat film. Hasil yang di dapatkan pada penelitian ini menunjukan nilai accuracy 65,57%, precision 81,20% dan recall 66,78 % dan berdasarkan analisa yang didapat menggunakan data set yang di dapat dari situs internet https://www.kaggle.com menunjukan bahwa mayoritas prediksi rating film "rendah".

Kata kunci: *analisa, data mining, film, naïve bayes*

PENDAHULUAN

Setiap bentuk kesenian, meliputi seni musik, seni tari, seni sastra, seni rupa ataupun seni peran perlu sebuah apresiasi dari penikmatnya masing-masing. Secara umum, apresiasi seni mempunyai makna penghargaan terhadap kehadiran sebuah karya seni, sebuah karya seni mengalami suatu perkembangan dari tahun ke tahun sehingga pada akhirnya tercipta sebuah perpaduan yang imbang dan juga harmonis antara seni sastra, seni musik, seni peran dan juga komedi yang dibungkus dalam bentuk film. Film adalah sarana baru yang dipergunakan untuk menyebarkan suatu hiburan yang telah menjadi kebiasaan terdahulu, dan menyajikan cerita peristiwa, musik, drama, lawak, dan sajian teknis lainnya kepada masyarakat umum (Mudjiono, 2011).

Ada beberapa alasan khusus mengenai masyarakat menyukai film, yaitu karena di dalam film terdapat unsur sebuah usaha masyarakat untuk mencari kesenangan serta mengisi waktu luang mereka, karena film terlihat hidup dan menarik perhatian masyarakat luas. Hal seperti ini bisa digunakan sebagai tujuan utama bagi produksi film agar mendapatkan suatu film yang bagus yakni yang dibungkus dalam berbagai cerita yang bisa membuat masyarakat tertarik, serta tak lupa juga diselipkan nilai-nilai yang bisa memperkaya hati agar disajikan kepada masyarakat luas sebagai panutan untuk hal-hal yang terdapat di dunia ini dengan pemahaman yang lebih *fresh*. Jadi film dapat dianggap sebagai wadah utuk mengekspresikan dan menggambarkan tentang kehidupan sehari-hari(Mudjiono, 2011). Untuk itu dibutuhkan analisa agar dapat mengetahui minat penikmat film dengan cara menganalisis rating suatu film menggunakan teknik data mining yaitu dengan menganalisis data dari pendapat yang berbeda dan merangkumnya untuk memperoleh suatu informasi yang bermanfaat. Informasi yang didapatkan dari hasil data mining bisa digunakan untuk meningkatkan pendapatan atau mengurangi biaya produksi(Sharma, Singh, & Singh, 2015). Pengumpulan data perlu dilakukan terlebih dahulu, biasanya data yang diperoleh bersifat *big data* yang berarti pengumpulan data dari berbagai macam sumber yang *relevan* (Utmal & Pandey, 2015).
Metode yang digunakan adalah *naïve bayes* yaitu metode yang mempunyai perhitungan matematik dasar yang sangat kuat serta

dalam efisiensi klasifikasinya juga stabil, namun kekurangannya adalah parameter model *naïve bayes* perlu diperkirakan dan kurang peka terhadap data yang sudah hilang. Model *naïve bayes* memiliki tingkat kesalahan yang sangat minimum jika dibandingkan dengan algoritma klasifikasi lainnya (Liu, Tian, Liu, Jiang, & Li, 2016). Metode *naïve bayes* ini merupakan salah satu metode yang populer untuk pengkategorian teks dengan frekuensi kata sebagai fitur. Hal ini dapat disimpulkan bahwa fitur-fitur yang independen dapat dibuktikan dalam algoritma klasifikasi menjadi lebih efektif (Chandrasekar & Qian, 2016).

TINJAUAN PUSTAKA

Telaah Penelitian

Dalam penelitian sebelumnya yang berhubungan dengan klasifikasi suatu dataset, untuk dijadikan sebagai bahan masukan penelitian diuraikan sebagai berikut:

1. Data mining sering disebut *Knowledge Discovery in Database* (KDD). Data mining biasanya digunakan untuk memperbaiki pengambilan keputusan di masa yang akan datang berdasarkan informasi yang diperoleh dari masa lalu. Misalnya untuk prediksi, estimasi, assosiasi, clustering, dan deskripsi. Sekumpulan data yang ada di laboratorium klinik belum difungsikan secara efektif dan hanya di fungsikan sebagai arsip untuk riwayat penyakit pasien. Jantung merupakan pembunuh nomor satu di dunia. Kurangnya aliran darah dan oksigen ke jantung bisa menyebabkan penyakit jantung. Pada penelitian ini akan membandingkan algoritma klasifikasi data mining Naive Bayes Berbasis PSO untuk deteksi penyakit jantung. Particle Swarm Optimization (PSO) merupakan inisiasi oleh sebuah populasi solusi acak dan selanjutnya mencari titik optimum dengan cara meng-update tiap hasil pembangkitan. Pendekatan yang digunakan lebih sistematis matematika untuk menemukan solusi. Pada eksperimen awal dihasilkan akurasi untuk algoritma naive bayes sebesar 82.14% dengan nilai area under cover (AUC) 0.686 dengan kategori “poor classification”. Pada eksperimen kedua dengan menggunakan algoritma naive bayes berbasis PSO menjadi 92.86% dan nilai AUC 0.839 dengan kategori “good classification”. Pada

eksperimen kedua terbukti bahwa dengan penambahan optimasi dapat meningkatkan nilai akurasi. Penelitian masih perlu dilakukan penelitian dengan menggunakan data yang lebih banyak dan menggunakan metode data mining yang lain.Menurut (Widiastuti, Santosa, & Supriyanto, 2014)

2. Utmal & Pandey (2015) mengemukakan penelitian dan industri saat ini bertujuan untuk bekerja dengan kumpulan data besar. skalabilitas tinggi dan toleransi kesalahan telah disediakan oleh beberapa solusi analitis data terdistribusi seperti Hadoop tetapi mereka tidak terlalu banyak user friendly dan fungsi mereka dapat dimanfaatkan sepenuhnya oleh pengembang. Hadoop merupakan framework open source berbasis Java di bawah lisensi Apache untuk mensupport aplikasi yang jalan pada Big Data. Hadoop berjalan pada lingkungan yang menyediakan storage dan komputasi secara terdistribusi ke kluster-kluster dari komputer/node. Tulisan ini mengusulkan konsep Radoop yang merupakan perpanjangan dari RapidMiner alat data mining dengan fungsi *Hadoop*. Kesimpulan kami di dalam makalah ini didasarkan pada kenyataan bahwa Radoop adalah alat yang dibuat untuk menangani analisis data yang besar dan bertemu baik dengan tuntutan yang semakin meningkat dari ukuran data.

3. Film memiliki nilai seni tersendiri, karena film tercipta sebagai sebuah karya dari tenaga-tenaga kreatif yang profesional di bidangnya. Film sebagai benda seni sebaiknya dinilai dengan secara artistik bukan rasional. Studi perfilman boleh dikatakan bidang studi yang relatif baru dan tidak sebanding dengan proses evolusi teknologinya. Semiotika merupakan suatu studi ilmu atau metode analisis untuk mengkaji tanda dalam suatu konteks skenario, gambar, teks, dan adegan di film menjadi sesuatu yang dapat dimaknai. Memaknai berarti bahwa obyek-obyek tidak hanya membawa informasi, dalam hal ini obyek-obyek itu hendak berkomunikasi, tetapi juga mengkonstitusi sistem terstruktur dari tanda yang digunakan dalam film tersebut (Mudjiono, 2011)

LANDASAN TEORI

Data Mining

Data Mining adalah penganalisisan data dari pendapat yang berbeda dan merangkumnya untuk memperoleh suatu informasi yang bermanfaat. Informasi yang didapatkan dari hasil data mining bisa digunakan untuk meningkatkan pendapatan atau mengurangi biaya produksi (Sharma et al., 2015) sedangkan menurut Utmal & Pandey (2015) Data Mining adalah Pengumpulan data perlu dilakukan terlebih dahulu, biasanya data yang diperoleh bersifat big data yang berarti pengumpulan data dari berbagai macam sumber yang relevan.

Naïve Bayes

Metode Naïve Bayes merupakan metode yang mempunyai perhitungan matematik dasar yang sangat kuat serta dalam efisiensi klasifikasinya juga stabil, namun kekurangannya adalah parameter model naïve bayes perlu diperkirakan dan kurang peka terhadap data yang sudah hilang. Model naïve bayes memiliki tingkat kesalahan yang sangat minimum jika dibandingkan dengan algoritma klasifikasi lainnya (Liu et al., 2016) dan Chandrasekar & Qian (2016) mengemukakan bahwa Metode naïve bayes ini merupakan salah satu metode yang populer untuk pengkategorian teks dengan frekuensi kata sebagai fitur. Hal ini dapat disimpulkan bahwa fitur-fitur yang independen dapat dibuktikan dalam algoritma klasifikasi menjadi lebih efektif.

Film

Setiap bentuk kesenian, meliputi seni musik, seni tari, seni sastra, seni rupa ataupun seni peran perlu sebuah apresiasi dari penikmatnya masing-masing. Secara umum, apresiasi seni mempunyai makna penghargaan terhadap kehadiran sebuah karya seni, sebuah karya seni mengalami suatu perkembangan dari tahun ke tahun sehingga pada akhirnya tercipta sebuah perpaduan yang imbang dan juga harmonis antara seni sastra, seni musik, seni peran dan juga komedi yang dibungkus dalam bentuk film. Film adalah sarana baru yang dipergunakan untuk menyebarkan suatu hiburan yang telah menjadi kebiasaan terdahulu, dan menyajikan cerita peristiwa, musik,

drama, lawak, dan sajian teknis lainnya kepada masyarakat umum (Mudjiono, 2011)

METODE PENELITIAN

Penentuan Obyek Observasi

Observasi ini bertujuan untuk memprediksi rating sebuah film. Observasi ini dipilih karena untuk tolak ukur sebuah rumah produksi film ketika nantinya akan membuat film serta mengetahui selera film yang sesuai dengan para penikmat film.

Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sampel dari berbagai situs di internet, yaitu dari situs https://www.kaggle.com/. Data tersebut digunakan sebagai data *training* dan juga data *testing*.

Data Training

Data *training* adalah data yang digunakan untuk perhitungan probabilitas dari data berdasarkan data pembelajaran yang dilakukan. Data *training* yang digunakan adalah data sampel yang di dapat dari situs di internet, yaitu dari situs https://www.kaggle.com/.

Data Testing

Data *testing* merupakan data yang akan atau sedang terjadi dan dipergunakan sebagai bahan uji yang sebelumnya sudah didapatkan pada data *training*. Data *testing* tersebut juga menggunakan data sampel yang diperoleh dari situs di internet, yaitu dari situs https://www.kaggle.com/.

Penentuan Atribut

Atribut-atribut yang digunakan untuk proses data mining ini mengacu pada tujuan penelitian. Ada dua jenis variabel yang ditentukan (Nugroho, & Setyawan, 2014),yaitu :

1. Variabel dependen (Y)

Variabel dependen (Y) merupakan variabel yang nilainya terikat, bisa disebut variabel terikat. Variabel Y yang digunakan yaitu imdb_score. Dengan melihat rating saja tidak cukup, karena yang memberi rating

adalah user-user dari IMDB, karena itu perlu dilihat juga angka user yang memberi rating (imdb_score).

2. Variabel independen (X)

Variabel independen (X) merupakan variabel yang nilainya tidak tergantung pada nilai dari variabel lainnya atau bisa disebut sebagai variabel bebas. Variabel tersebut nantinya akan sangat mempengaruhi hasil dari prediksi film karena rating suatu film berhubungan dengan variabel-variabel tersebut. Variabel X yang digunakan terdiri dari:

1) Variabel X1 Color
2) Variabel X2 Director_name
3) Variabel X3 Director_facebook_likes
4) Variabel X4 Duration
5) Variabel X5 Actor_1_name
6) Variabel X6 Actor_1_facebook_likes
7) Variabel X7 Actor_2_name
8) Variabel X8 Actor_2_facebook_likes
9) Variabel X9 Actor_3_name
10) Variabel X10 Actor_3_facebook_likes
11) Variabel X11 Gross
12) Variabel X12 Genres
13) Variabel X13 Movie_title
14) Variabel X14 Num_voted_users
15) Variabel X15 Cast_total_facebook
16) Variabel X16 Facenumber_in_poster
17) Variabel X17 Plot_Keywords
18) Variabel X18 Num_users_for_reviews
19) Variabel X19 Num_critic_for_reviews
20) Variabel X20 Language
21) Variabel X21 Country
22) Variabel X22 Content_Rating
23) Variabel X23 Budget
24) Variabel X24 Title_Year
25) Variabel X25 Movie_fb_Like

Data Cleaning

Pembersihan data perlu dilakukan supaya data yang digunakan valid sesuai kebutuhan. Sehingga dari nilai *class* data film dalam atribut

tidak terjadi ketidakkonsistenan data dalam pengujian. Nilai-nilai *class* yang terdapat dalam setiap atribut dijabarkan pada tabel 1.

Tabel 1. Nilai class pada setiap atribut

Nama Atribut	Notasi	Class
Color	X1	Color , Black and White
Director_Name	X2	Text
Director_Facebook_Likes	X3	Rendah ≤ 100, Sedang < 500, Tinggi ≥ 500
Duration	X4	Pendek ≤ 120, Panjang ≥ 120
Actor_1_Name	X5	Text
Actor_1_Facebook_Likes	X6	Rendah ≤ 1000, Sedang < 5000, Tinggi ≥ 5000
Actor_2_Name	X7	Text
Actor_2_Facebook_Likes	X8	Rendah ≤ 500, Sedang < 1000, Tinggi ≥ 1000
Actor_3_Name	X9	Text
Actor_3_Facebook_Likes	X10	Rendah ≤ 400, Sedang < 800, Tinggi ≥ 800
Gross	X11	Rendah ≤ 50000000, Sedang < 100000000, Tinggi ≥100000000
Genres	X12	Text
Movie_Title	X13	Text
Num_Voted_Users	X14	Rendah ≤ 100000, Sedang < 250000, Tinggi ≥ 250000
Cast_Total_Facebook_Likes	X15	Rendah ≤ 5000, Sedang < 30000, Tinggi ≥ 30000
Face_Number_In_Poster	X16	Rendah ≤ 1, Sedang < 3, Tinggi ≥ 3
Plot_Keywords	X17	Text
Num_Users_For_Reviews	X18	Rendah ≤ 300, Sedang < 700, Tinggi ≥ 700
Num_Critic_For_Reviews	X19	Rendah ≤ 150, Sedang < 300, Tinggi ≥ 300
Language	X20	Text

Nama Atribut	Notasi	Class
Country	X21	Text
Content_Rating	X22	Text
Budget	X23	Rendah ≤ 30000000, Tinggi ≥ 30000000
Title_Year	X24	Lama ≤ 2000, Baru > 2000
Movie_facebook_Likes	X25	Rendah ≤ 5000, Sedang < 10000, Tinggi ≥ 10000

Penggunaan Metode *Naïve Bayes*

Naïve Bayes adalah sebuah pengelompokan statistik yang bisa di dipakai untuk memprediksi probabilitas anggota suatu *class*. Naïve Bayes juga mempunyai akurasi dan kecepatan yang sangat kuat ketika diaplikasikan pada database dengan *big data* (Widiastuti et al., 2014). Berikut rumus *naive bayes* (Nugroho & Haryati, 2015) ditunjukkan pada persamaan 1.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P\ (X)} = P(X|Y)P(Y) \qquad (1)$$

Keterangan:

X : Data dengan *class* yang belum diketahui
Y : Hipotesis data yaitu suatu class spesifik
P(Y|X) : Probabilitas hipotesis berdasar kondisi X (*posteriori probability*)
P(Y) : Probabilitas hipotesisY (*prior probability*)
P(X|Y) : Probabilitas X saat kondisi hipotesis Y
P (X) : Probabilitas X

Pengujian *Accuracy, Precicion, Recall*

Diperlukan pengujian *Accuracy, Precicion, Recall* agar bisa membuktikan kinerja metode naïve bayes berikut rumus pengujian *Accuracy, Precicion, Recall* (Nugroho & Gunawan, 2016)

a. Accuracy

$$\frac{(TP+TN)}{(TP+FP+TN+FN)} \qquad (2)$$

b. Recall

$$\frac{TP}{TP+FN} \quad (3)$$

c. Precision

$$\frac{TP}{TP+FP} \quad (4)$$

Tabel 2. Keputusan yang didapatkan dari data *training* dan data *testing*

Clasification	**Predicted Class**	
Observed class	**Class = YES**	**Class = NO**
Class = YES	True Positive	False Positive
Class = NO	False Negative	True Negative

Sumber :(Agustina & Wijanarto, 2016)

HASIL DAN PEMBAHASAN

Data yang dipakai adalah data set yang diambil dari situs https://www.kaggle.com/ yang akhirnya dipakai untuk data *training* dan data *testing*. Kemudian dilakukan *prepocessing* pada data tersebut dengan cara menerapkan metode *cleaning*. Lalu data tersebut melalui pengolahan, pengolahan data menggunakan *software rapid miner* yang nantinya bisa menghasilkan sebuah prediksi rating film. Tabel 3 menunjukkan potongan data sebelum tahap *preprocessing*. Sedangkan tabel 4 merupakan potongan data sesudah tahap *preprocessing*.

Tabel 3. Data sebelum *prepocessing*

color	director_ name	num_critic_ for_reviews	duration	director_ facebook_likes	actor_3_ facebook_likes	actor_2_name
Color	James Cameron	723	178	0	855	Joel David Moore
Color	Gore Verbinski	302	169	563	1000	Orlando Bloom
Color	Sam Mendes	602	148	0	161	Rory Kinnear
Color	Christopher Nolan	813	164	22000	23000	Christian Bale
Color	Andrew Stanton	462	132	475	530	Samantha Morton
Color	Sam Raimi	392	156	0	4000	James Franco
Color	Nathan Greno	324	100	15	284	Donna Murphy
Color	Joss Whedon	635	141	0	19000	Robert Downey Jr.
Color	David Yates	375	153	282	10000	Daniel Radcliffe
Color	Zack Snyder	673	183	0	2000	Lauren Cohan

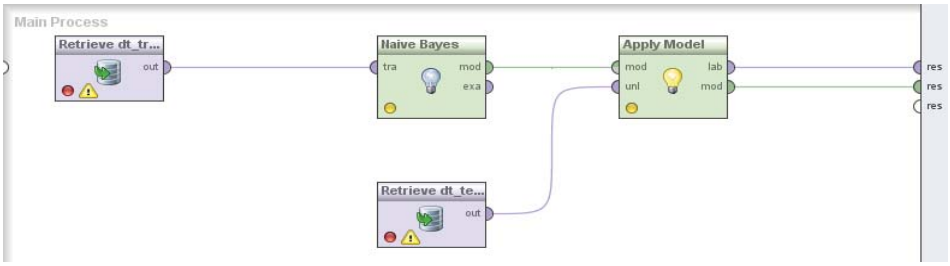
Tabel 4. Data sesudah *prepocessing*

color	director_name	director_ facebook_like	Duration	actor_1_name	actor_1_ facebook_likes	actor_2_name
Color	James Cameron	rendah	Panjang	CCH Pounder	rendah	Joel David Moore
Color	Gore Verbinski	tinggi	Panjang	Johnny Depp	tinggi	Orlando Bloom
Color	Sam Mendes	rendah	Panjang	Christoph Waltz	tinggi	Rory Kinnear
Color	Christopher Nolan	tinggi	Panjang	Tom Hardy	tinggi	Christian Bale
Color	Andrew Stanton	sedang	Panjang	Daryl Sabara	rendah	Samantha Morton
Color	Sam Raimi	rendah	panjang	J.K. Simmons	tinggi	James Franco
Color	Nathan Greno	rendah	pendek	Brad Garrett	rendah	Donna Murphy
Color	Joss Whedon	rendah	panjang	Chris Hemsworth	tinggi	Robert Downey Jr.
Color	David Yates	sedang	panjang	Alan Rickman	tinggi	Daniel Radcliffe
Color	Zack Snyder	rendah	panjang	Henry Cavill	tinggi	Lauren Cohan

Implementasi *Software Rapid Miner*

Penggunaan software ini bisa mengimpor sebuah informasi yang terdapat dari berbagai macam sumber database untuk diperiksa dan dianalisa didalam sebuah aplikasi. Rapid miner sebagai solusi untuk memprediksi dan menganalisa komputasi statistik (Utmal & Pandey, 2015)

Gambar 1 merupakan rancangan proses klasifikasi naïve bayes yang dilakukan menggunakan aplikasi *rapid miner*:



Gambar 1. Klasifikasi *Naïve Bayes*

SimpleDistribution

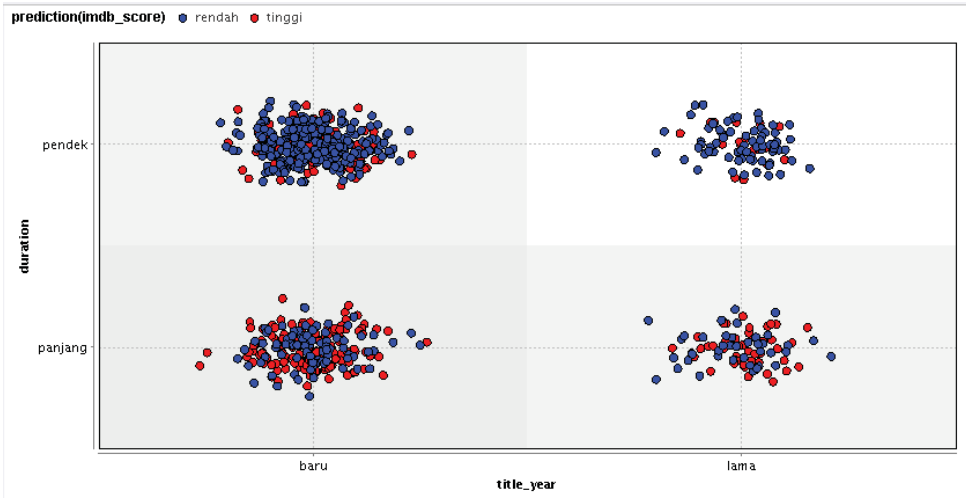
```
Distribution model for label attribute imdb_score


Class rendah (0.707)
25 distributions

Class tinggi (0.293)
25 distributions
```

Gambar 2. Hasil *Naïve Bayes* pada text view

Hasil pada gambar 2 menunjukan model distribusi *Naïve Bayes*. Pada hasil *naïve bayes* bisa dilihat bahwa model distribusi nilai *class* "RENDAH" sebanyak 0.707, sedangkan *class* "TINGGI" sebanyak 0.293.



Gambar 3. Hasil *Naïve Bayes* pada plot view

Berdasarkan gambar 3, penentuan rating film (*imdb_score*) tinggi apabila tahun film (*title_year*) baru serta *duration* film panjang.

Pengujian

Hasil klasifikasi yang didapat tersebut kemudian dihitung nilai performanya berdasarkan *accuracy, precision, recall.* Tabel 7 adalah hasil perhitungan tingkat performa algoritma naïve bayes.

Tabel 7. Tingkat *Accuracy, Precision, Recall* pada *Naïve bayes*

Accuracy = 65,57%			
	True Tinggi	**True Rendah**	**Class Precision**
Prediksi Tinggi	1417	328	81,20%
Prediksi Rendah	705	550	43,82%
Class Recall	66,78%	62,64%	

Berdasarkan tabel 7 nilai *accuracy* 65,57%, *precision* 81,20% dan *recall* 66,78 %. Pengujian tersebut juga bisa dihitung dengan menggunakan rumus persamaan 2, persamaan 3, persamaan 4.

Tabel 8. Nilai Keputusan dari data *training* dan data *testing*

Clasification	**Predicted Class**	
Observed class	**Class = YES**	**Class = NO**
Class = YES	1417	328
Class = NO	705	550

a. Accuracy

Berdasarkan rumus *Accuracy* pada persamaan 2 maka dapat dihitung

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)}$$

$$= \frac{1417+550}{1417+328+550+705}$$

= 0,6556 x 100%

= 65,56%

b. Recall

Berdasarkan rumus *Recall* pada persamaan 3 maka dapat dihitung:

$$Recall = \frac{TP}{TP+FN}$$

$$= \frac{1417}{1417+705}$$

= 0,6677 x 100%

= 66,77%

c. Precision

Berdasarkan rumus *Precision* pada persamaan 3 maka dapat dihitung:

$$Precision = \frac{TP}{TP+FP}$$

$$= \frac{1417}{1417+328}$$

= 0,8120 x 100%

= 81,20%

Perhitungan *Naïve Bayes* (HMAP)

Perhitungan HMAP data set yang diambil dari situs internet https://www.kaggle.com/ yaitu keseluruhan ada 3806 data dan dari variabel prediksi yaitu imdb_score ada dua *class* yaitu TINGGI dan

RENDAH, jumlah data pada *class* tinggi adalah 1098 dan pada *class* rendah 2708. Perhitungan HMAP menggunakan *sample* beberapa atribut X, yakni 7 (tujuh) atribut.

P (Y = TINGGI) = 1098/3806 = 0,27
P (Y = RENDAH) = 2708/3806 = 0,71

1. Fakta 1(merupakan atribut-atribut yang mayoritas *class* tinggi ataupun panjang).
   P(X1=Color | Y=Tinggi) = 1035/1098 = 0,94
   P(X1=Color | Y=Rendah) = 2646/2708 = 0,97
   P(X3=Tinggi | Y=Tinggi) = 229/1098= 0,20
   P(X3=Tinggi | Y=Rendah) = 258/2708 = 0,09
   P(X4=Panjang | Y=Tinggi) = 448/1098 = 0,40
   P(X4=Panjang | Y=Rendah) = 460/2708 = 0,16
   P(X6=Tinggi | Y=Tinggi) = 511/1098 = 0,46
   P(X6=Tinggi | Y=Rendah) = 964/2708 = 0,35
   P(X8=Tinggi | Y=Tinggi) = 283/1098 = 0,25
   P(X8=Tinggi | Y=Rendah) = 604/2708 = 0,22
   P(X10=Tinggi | Y=Tinggi) = 203/1098 = 0,18
   P(X10=Tinggi | Y=Rendah) = 481/2708 = 0,31
   P(X11=Tinggi | Y=Tinggi) = 252/1098 = 0,22
   P(X11=Tinggi | Y=Rendah) = 349/2708 = 0,11
   Y = Tinggi
   =0,94x0,20x0,40x0,46x0,25x0,18x0,22
   =0,00034
   Y = Rendah
   =0,97x0,09x0,16x0,35x0,22x0,31x0,11
   =0,000021

2. Fakta 2 (merupakan atribut-atribut yang mayoritas *class* sedang).
   P(X3 = Sedang | Y=Tinggi) = 321/1098 = 0,29
   P(X3 = Sedang | Y=Rendah) = 721/2708 = 0,26
   P(X6 = Sedang | Y=Tinggi) = 79/1098 = 0,07
   P(X6 = Sedang | Y=Rendah) = 274/2708 = 0,10
   P(X8 = Sedang | Y=Tinggi) = 419/1098 = 0,38
   P(X8 = Sedang | Y=Rendah) = 1207/2708 = 0,44
   P(X10 = Sedang | Y=Tinggi) = 352/1098 = 0,32
   P(X10 = Sedang | Y=Rendah) = 996/2708 = 0,36

P(X11 = Sedang | Y=Tinggi) = 188/1098 = 0,17
P(X11 = Sedang | Y=Rendah) = 494/2708 = 0,18
P(X14 = Sedang | Y=Tinggi) = 318/1098 = 0,28
P(X14 = Sedang | Y=Rendah) = 444/2708 = 0,16
P(X15 = Sedang | Y=Tinggi) = 404/1098 = 0,36
P(X15 = Sedang | Y=Rendah) = 928/2708 = 0,34
Y= Tinggi
=0,29x0,07x0,38x0,32x0,17x0,28x0,36
= 0,000422
Y= Rendah
=0,26x0,10x0,44x0,36x0,18x0,16x0,34
= 0,00004032

3. Fakta 3 (merupakan atribut-atribut yang mayoritas *class*nya rendah ataupun pendek)
   P(X1=BNW | Y=Tinggi) = 63/1098 = 0,05
   P(X1=BNW | Y=Rendah) = 62/2708 = 0,02
   P(X3=Rendah | Y=Tinggi) = 548/1098 = 0,49
   P(X3=Rendah | Y=Rendah) =1719/2708 = 0,63
   P(X4=Pendek | Y=Tinggi) = 650/1098 = 0,59
   P(X4=Pendek | Y=Rendah) = 2248/2708 = 0,83
   P(X6=Rendah | Y=Tinggi) = 508/1098 = 0,46
   P(X6=Rendah | Y=Rendah) = 1470/2708 = 0,54
   P(X8=Rendah | Y=Tinggi) = 396/1098 = 0,36
   P(X8=Rendah | Y=Rendah) = 897/2708 = 0,33
   P(X10=Rendah | Y=Tinggi) = 543/1098 = 0,49
   P(X10=Rendah | Y=Rendah) = 1231/2708 = 0,45
   P(X11=Rendah | Y=Tinggi) = 658/1098 = 0,59
   P(X111=Rendah | Y=Rendah) = 1865/2708 = 0,68
   Y=Tinggi
   =0,05x0,49x0,59x0,46x0,36x0,49x0,59
   = 0,00069203
   Y=Rendah
   =0,02x0,63x0,83x0,54x0,33x0,45x0,68
   = 0,000570

Jadi pada saat menggunakan aplikasi *rapid miner* dan perhitungan HMAP manual menyatakan bahwa rating film pada data

yang digunakan cenderung **rendah.** Karena jika dibandingkan fakta 1, fakta 2, dan fakta 3 yang mempunyai nilai terbesar adalah pada fakta 3 yang didalam fakta 3 tersebut terdapat atribut-atribut yang *class* nya rendah, yakni perbandingannya pada Y = Tinngi (fakta 1 = 0,00034, fakta 2 = 0,000422 dan fakta 3 = 0,00069203) dan pada kelas Y = Rendah (fakta 1 = 0,000021, fakta 2 = 0,00004032 dan fakta 3 = 0,000570)

KESIMPULAN

Berdasarkan hasil penelitian, maka penulis dapat menarik kesimpulan bahwa Naïve Bayes adalah salah satu klasifikasi berjenis teks, contohnya adalah prediksi rating film ini. Naïve Bayes ini sangat familiar jika dipakai untuk pengklasifikasian teks dan mempunyai performa yang bagus pada banyak domain. Hasil penelitian menunjukkan bahwa hasil prediksi rating film menggunakan metode naïve bayes memiliki *accuracy* 65,56 %, *precision* 81,20%, dan *recall* 66,77%, Berdasarkan analisa yang didapat menggunakan data set dari situs https://www.kaggle.com/ dengan menggunakan data *testing* sebanyak 806 data menunjukan bahwa mayoritas prediksi rating film **RENDAH.**

DAFTAR PUSTAKA

Agustina, D. M., & Wijanarto. (2016). Analisis Perbandingan Algoritma ID3 Dan C4 . 5 Untuk Klasifikasi Penerima Hibah Pemasangan Air Minum, *1*(3), 234–244.

Chandrasekar, P., & Qian, K. (2016). The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier. *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, *2*, 618–619. http://doi.org/10.1109/COMPSAC.2016.205

Liu, J., Tian, Z., Liu, P., Jiang, J., & Li, Z. (2016). An Approach of Semantic Web Service Classification Based on Naive Bayes. *2016 IEEE International Conference on Services Computing An*. http://doi.org/10.1109/SCC.2016.53

Mudjiono, Y. (2011). Kajian Semiotika Dalam Film. *Jurnal Ilmu Komunikasi, Vol. 1, No.1, April 2011 ISSN: 2088-981X KAJIAN*, *1*(1), 123.

Nugroho, Y. S., & Gunawan, D. (2016). Decision Tree Induction for Classifying the Cholesterol Levels. *The 2nd International Conference on Science, Technology, and Humanity*, 231–240.

Nugroho, Y. S., & Haryati, S. N. (2015). Klasifikasi dan Klastering Penjurusan Siswa SMA Negeri 3 Boyolali. *Klasifikasi Dan Klastering Penjurusan Siswa SMA Negeri 3 Boyolali*, *1*(1), 3–8.

Nugroho, Y. S., & Setyawan. (2014). Klasifikasi Masa Studi Mahasiswa Fakultas Komunikasi Dan Informatika Universitas Muhammadiyah Surakarta Menggunakan Algoritma C4.5. *KomuniTi, Vol. VI, No. 1 Maret 2014*, *VI*(1), 84–91.

Sharma, P., Singh, D., & Singh, A. (2015). Classification Algorithms on a Large Continuous Random Dataset Using Rapid. *IEEE Sponsored 2nd International Conference On Electronics And Communication System (ICECS 2015)*, (Icecs), 704–709.

Utmal, M., & Pandey, R. K. (2015). Taxonomy on the Integration of Hadoop and Rapid Miner for Big Data Analytics. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 890–893). http://doi.org/10.1109/CICN.2015.175

Widiastuti, N. A., Santosa, S., & Supriyanto, C. (2014). Algoritma Klasifikasi Data Mining Naïve Bayes Berbasis Particle Swarm Particle Swarm Optimization Untuk Deteksi Penyakit Jantung. *Jurnal Pseudocode*, *1*, 11–14.