JURNAL ILMIAH TEKNOLOGI INFORMASI DAN KOMUNIKASI (JTIK) VOL 14, No.2, September 2023, pp. 218 - 227
p-ISSN : 2087-0868 (media cetak)
e-ISSN : 2598-9707 (media online)
http://ejurnal.provisi.ac.id/index.php/JTIKP

THE USE OF NAÏVE BAYES ALGORITHM IN FORECASTING THE FURTHER STUDY OF HIGH SCHOOL STUDENT

Siti Nur Amalia[1], Maulana Wildan Rifaldi[2], Mega Aprilia Fajriati[3], Rona Nisa Sofia Amriza[4]
[1,2,3,4] Sistem Informasi, IT Telkom Purwokerto - Banyumas, Jawa Tengah, Indonesia e-mail: 18103091@ittelkom-pwt.ac.id[1], 19103028@ittelkom-pwt.ac.id[2], 18103050@ittelkom-pwt.ac.id[3], rona@ittelkom-pwt.ac.id[4]

ARTICLE INFO

Article history:

Received : 17 – Februari - 2023
Received in revised form : 4 – Maret - 2023
Accepted : 5 – April - 2023
Available online : 1 – September - 2023

ABSTRACT

Schools with a large number of students who continue their studies to college will be the main choice as secondary schools from the elementary level. Therefore, increasing the number of students continuing their studies is very important to meet the competition. To get around this, schools can predict the continuation of high school/vocational high school students' studies to college. The goal is that the percentage of prediction results can be used as a reference for improving the quality of education services in schools. In making this prediction, the Naïve Bayes method or algorithm is used. In this case, the Naïve Bayes algorithm is a classification method with a probability and statistical approach that is suitable for predicting the continuation of high school/vocational high school students' studies to college. The prediction result of continuing study to university using Naïve Bayes on test data has an accuracy of 0.740.

Keywords: Naïve Bayes, prediction, data mining.

1. PENDAHULUAN

Perkembangan yang begitu pesat di bidang pendidikan, membuat semakin ketatnya kompetisi diantara sekolah. Sekolah dengan banyak jumlah peserta didik yang melanjutkan studi ke perguruan tinggi akan menjadi pilihan utama sebagai sekolah lanjutan dari tingkat dasar. Maka dari itu, meningkatkan jumlah siswa lanjut studi menjadi sangat penting untuk memenuhi kompetisi tersebut. Untuk menyiasati hal ini, sekolah dapat melakukan prediksi kelanjutan studi siswa SMA/SMK ke perguruan tinggi. Tujuannya yakni agar prosentase hasil prediksi dapat dijadikan sebagai acuan untuk perbaikan mutu pelayanan pendidikan di sekolah.[1] Data mining merupakan metode pembelajaran yang mencakup beberapa teknik yang telah di teliti dan dikembangkan baik untuk kebutuhan industri, pendidikan, komersial maupun kebutuhan ilmiah. Data mining memiliki beberapa teknik yang berbeda untuk menyelesaikan permasalahan yang berbeda pula, salah satunya yaitu teknik klasifikasi. Tujuan dari teknik klasifikasi adalah untuk membagi objek ke dalam kategori yang disebut kelas. Data yang digunakan pada penelitian ini adalah sample dari dataset klasifikasi pada www.kaggle.com. Dalam melakukan prediksi ini digunakan teknik klasifikasi dengan algoritma Naïve Bayes. Dalam hal ini algoritma Naïve Bayes merupakan metode klasifikasi dengan pendekatan probabilitas dan statistik yang cocok dalam membuat prediksi kelanjutan studi siswa SMA/SMK ke perguruan tinggi, karena memiliki rumus yang sederhana dan mudah untuk diaplikasikan[2].

Received 17 – Februari 2023; Revised 4 – Maret - 2023; Accepted 5 – April - 2023

p-ISSN : 2087-0868 e-ISSN : 2598-9707

2. TINJAUAN PUSTAKA

2.1. Dasar Teori

2.1.1. Data Mining

Data mining atau penambangan data adalah serangkaian proses untuk mendapatkan pengetahuan atau pola dari kumpulan data. Penambangan data akan memecahkan masalah dengan menganalisis data yang sudah ada dalam database. Penambangan data sering juga disebut Knowledge Discovery in Databases (KDD) yaitu suatu aktivitas yang meliputi pengumpulan, penggunaan data historis untuk menemukan pola reguler, pola hubungan dalam kumpulan data yang besar[3]. Sebagai suatu rangkaian proses, Data Mining dapat dibagi menjadi beberapa tahap proses. Tahap-tahap Data Mining antara lain sebagai berikut [4]:

a. Pembersihan Data (Data Cleaning)
   Yaitu proses menghilangkan noise dan data yang tidak konsisten atau tidak relevan, misalkan terdapat data yang tidak valid (null) atau data yang kosong sehingga tidak diperlukan. Sehingga data otomatis akan dibersihkan dalam artian dihapus karena keberadaannya bisa mengurangi mutu atau akurasi dari hasil data mining nantinya.
b. Integrasi Data (Data Integration)
   Yaitu penggabungan data dari berbagai database ke dalam suatu database baru. Dalam tahap ini dilakukan pada atribut-atribut yang dapat mengidentifikasi suatu entitas-entitas yang unik seperti nama,jenis produk, atribut, nomor pelanggan. Bila terjadi kesalahan pada tahap integrasi data bisa menghasilkan hasil yang menyimpang dan dapat menyesatkan pengambilan aksi pada tahap selanjutnya.
c. Seleksi data (Data Selection)
   Data yang ada pada database seringkali tidak semuanya dipakai, maka dari itu hanya data yang sesuai untuk dianalisis yang akan diambil dari database.
d. Transformasi data (*Data Transformation)*
   Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam Data Mining. Beberapa teknik data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa teknik standar seperti analisis asosiasi dan klastering hanya bisa menerima input data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut binning. Disini juga dilakukan pemilihan data yang diperlukan oleh teknik data mining yang dipakai. Transformasi dan pemilihan data ini juga menentukan kualitas dari hasil data mining nantinya karena ada beberapa karakteristik dari teknik-teknik data mining tertentu yang tergantung pada tahapan ini.
e. Proses Mining
   Yaitu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data. Beberapa metode yang dapat digunakan berdasarkan pengelompokan data mining, pada proses ini, mulai dari teknik, metode, hingga algoritma yang digunakan untuk data mining sangat bervariasi.
f. Evaluasi Pola (*Pattern Evaluation)*
   Yaitu mengidentifikasi pola-pola menarik ke dalam knowledge based yang ditemukan. evaluasi pola yang ditemukan (proses interprestasi pola menjadi pengetahuan yang dapat digunakan untuk mendukung pengambilan keputusan)
g. *Knowledge Presentation*
   Yaitu visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Pada tahap terakhir pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan.[4]

2.1.2. Naïve Bayes

Naïve Bayes merupakan sebauah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Teorema Bayes mengasumsikan semua atribut independen atau tidak saling ketergantungan antara nilai pada variabel kelas. Definisi lain menyebutkan Naïve Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuan Inggris Thomas bayes, yaitu memprediksi peuang di masa depan berdasarkan pengalaman dimasa sebelumnya. Naïve Bayes didasarkan pada asumsipenyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Keuntungan penggunaan Naïve Bayes adalah bahwa metode ini hanya membutuhkan jumlah data latih yang kecil untuk mengetahui estimasi parameter yang diperlukan dalam proses pengklasisfikasian[4].

p-ISSN : 2087-0868 e-ISSN : 2598-9707

Persamaan Metode Naïve Bayes :

$$P(H|X) = \frac{P(X \mid H).P(H)}{P(X)}$$

Karena asumsi atribut tidak saling terkait, maka :

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

Bila $P(X|C_i)$ dapat diketahui melalui perhitungan di atas, maka klas (label) dari data sampel X adalah klas (label) yang memiliki $P(X|C_i) * P(Ci)$ maksimum[5].

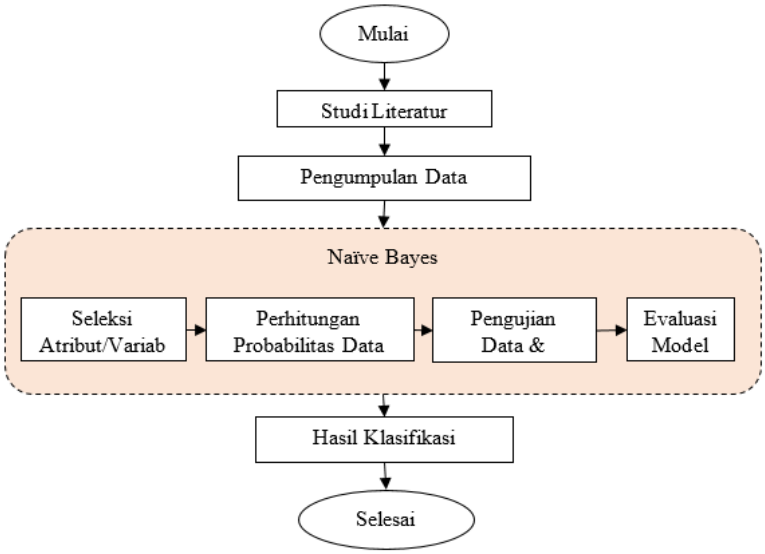
2.2. Penelitian Sebelumnya

Penelitian ini berjudul "Penerapan Algoritma Naïve Bayes pada Prediksi Kelanjutan Studi Siswa SMA/SMK". Penelitian ini didasari oleh beberapa penelitian yang pernah dilakukan sebelumnya, antara lain: Penelitian [1] menggunakan metode Naïve Bayes untuk memprediksi kelanjutan studi siswa ke perguruan tinggi. Karena dengan mengetahui jumlah siswa yang melanjutkan atau tidak, dapat menjadi acuan untuk meningkatakan pelayanan pendidikan pada sekolah yang bersangkutan. Pada penelitian ini digunakan K-fold validation dan confusion matrix, dan mendapatkan nilai prediksi sebesar 86.53%[1]. Selanjutnya penelitian [4] menggunakan metode klasifikasi Naïve Bayes untuk memprediksi besarnya penggunaan listrik rumah tangga. Penerapan Naïve Bayes diharapkan mampu untuk memprediksi besarnya penggunaan listrik rumah tangga. Berdasarkan data rumah tangga yang ijadikan data training, metode Naïve Bayes mengklasifikasika 47 data dari 60 data yang diuji. Dan hasil prediksi penggunaan listrik rumah tinggi mencapai keakuratan 78,3333%. Kemudian penelitian [3] menggunakan metode Naïve Bayes untuk memprediksi kelulusan mahasiswa. Diharapkan tinggi nya jumlah mahasiswa yang tidak lulus tepat waktu dapat diminimalkan dengan melakukan prediksi dini. Hasil penelitian ini membuktikan bahwa hasil prediksi kelulusan mahasiswa menggunakan Naïve Bayes Classification menghasilkan akurasi 73.725%, precision 0.742, recall 0.736 dan F-measure 0.735. Selanjutnya penelitian [6] menggunakan metode Naïve Bayes untuk memprediksi usia kelahiran bayi. Hal ini didasarkan pada banyaknya kematian bayi akibat usia kelahiran yang kurang mencukupi atau yang lewat waktu. Penelitian ini bertujuan untuk membuat aplikasi prediksi yang nantinya akan dapat membantu pasien dalam mengetahui usia kelahirannya dan mengantisipasi hal yang tidak diinginkan kedepannya. Akurasi prediksi yang dihasilkan pada penelitian ini mencapai 78.69%, precision 70.14% dan recall 63.64%.

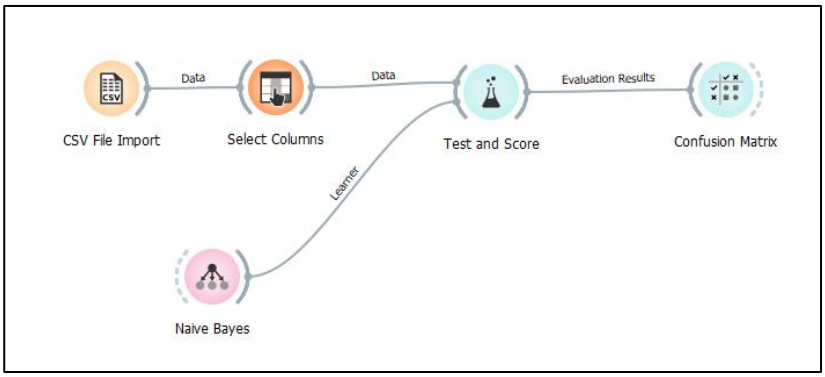
3. METODOLOGI PENELITIAN

Penelitian ini didasarkan pada metode eksperimental terhadap data. Tahapan penelitian ini terlihat pada Gambar 1. Tahapan pertama yaitu studi literatur mengenai teori seputar data mining dan Naïve Bayes yang bersumber dari jurnal-jurnal penelitian sebelumnya. Tahap selanjutnya yakni pengumpulan data. Data yang diambil untuk penelitian ini merupakan dataset klasifikasi pada pada www.kaggle.com. Data terdiri dari data siswa berasal dari SMA/SMK. Pengumpulan data ini dimulai dari data selection, cleaning, hingga splitting data. Tahap yang ketiga yaitu pemrosesan data dengan Naïve Bayes. Tedapat 4 tahapan pada Naïve Bayes ini, yaitu Seleksi atribut, Perhitungan probabilitas, Pengujian data, dan Evaluasi Model. Pada seleksi atribut, dilakukan penentuan variabel yang akan digunakan untuk mengklasifikasikan data. Selanjutnya dilakukan perhitungan probabilitas P(Xk|Ci) untuk setiap class data training. Kemudian data testing diuji berdasarkan perhitungan probabilitas yang telah dilakukan. Selanjutnya Evaluasi model dilakukan untuk mengetahui confusion mattix yang dihasilkan. Pada Gambar 2, menunjukkan proses Evaluai Model menggunakan aplikasi Orange. Pertama-tama, mengimport file dataset dengan format csv yang akan dievaluasi. Kemudian menentukan kolom-kolom pada dataset yang akan dievaluasi. Selanjutnya menghubungkan Confusion Matrix dengan metode Naïve Bayes menggunakan Test and Score.

p-ISSN : 2087-0868 e-ISSN : 2598-9707



Gambar 1. Metodologi Penelitian



Gambar 2. Model Data Mining

4. HASIL DAN PEMBAHASAN

4.1. Data Selection

Data yang akan digunakan pada proses data mining ini merupakan dataset pada Kaggle yang berisi data sintetis untuk proyek perguruan tinggi, yaitu data siswa berasal dari SMA/SMK. Data ini bertujuan untuk memprediksi apakah siswa SMA/SMK tersebut akan melanjutkan ke perguruan tinggi atau tidak. Dari dataset tersebut diambil 50 sampel data sebagai bahan perhitungan probabilitas.

4.2. Data Prepocessing/Cleaning

Pada tahap ini data diperiksa, jika ditemukan duplikasi data, data inkonsisten, data yang salah atau data yang kosong maka data perlu di sesuaikan dengan cara : diperbaiki, ditambahkan, dilengkapi atau dihapus. Sehingga data bersih dan siap untuk diproses. Data yang akan diproses terdiri dari 50 data dengan 11 kolom (Type school, school accreditation, gender, interest, residence, parent age, parent salary, house area, average grades, parent was in college, in college). Tidak ada data yang kosong pada dataset tersebut, hanya terdapat beberapa data yang masih perlu disesuaikan agar lebih konsisten. Data yang disesuaikan yaitu data pada kolom Parent age, House area,Parent salary, dan Average grades. Data pada kolom tersebut diklasifikasikan agar lebih rapi dan mudah dalam melakukan perhitungan maupun pengujian. Sampel data sebelum prepocessing atau cleaning tersaji pada Tabel 1, sedangkan sampel data setelah preprocessing disajikan pada Tabel 2.

p-ISSN : 2087-0868 e-ISSN : 2598-9707

Tabel 1. Sampel Data Sebelum Prepocessing Atau Cleaning

Type school	school accre-dittation	gender	interest	reside nce	Parent age	Parent salary	House area	Average grades	Parent Was In college	In college
Academic	A	Male	Less Interested	Urban	56	6950000	83	84,09	False	True
Academic	B	Female	Very Interested	Urban	50	6500000	80,6	87,43	False	True
Vocational	B	Male	Not Interested	Rural	49	6600000	78,2	82,12	True	True
Academic	A	Female	Uncertain	Urban	57	5250000	75,1	86,79	False	False
Vocational	B	Female	Less Interested	Rural	48	3770000	65,3	86,79	True	False

Tabel 2. Sampel Data Setelah Preprocessing Atau Cleaning

Type school	school accre-ditta-tion	gender	interest	reside nce	Parent age	Parent salary	House area	Average grades	Parent Was In college	In college
Academic	A	Male	Less Interested	Urban	> 50	>5,5 Juta	> 70	<= 86,5	False	True
Academic	B	Female	Very Interested	Urban	> 50	>5,5 Juta	> 70	> 86,5	False	True
Vocational	B	Male	Not Interested	Rural	<=50	>5,5 Juta	> 70	<= 86,5	True	True
Academic	A	Female	Uncertain	Urban	> 50	>5,5 Juta	> 70	> 86,5	False	False
Vocational	B	Female	Less Interested	Rural	<=50	<= 5,5 Juta	<=70	> 86,5	True	False

4.3. Splitting Data

Klasifikasi Naïve Bayes memiliki dua tahap, yaitu tahap training dan tahap testing. Pada tahap training, Naïve Bayes akan mempelajari data terlebih dahulu agar dapat membentuk model probabilitas. Selanjutnya, model probabilitas yang telah dibentuk berdasarkan tahap training, akan diuji menggunakan sebuah data. Pengujian dilakukan untuk menguji metode yang digunakan, apakah dapat menghasilkan akurasi yang baik atau tidak. Tahap pengujian ini disebut tahap testing. Oleh karena itu, dataset yang ada dikelompokkan menjadi 2 (data training dan data testing) secara acak. Data training terdiri dari 50 data yang telah melalui preprocessing atau cleaning. Sedangkan data testing, terdiri dari 5 data yang diambil secara acak.

4.4. Naïve Bayes

Langkah-langkah perhitungan Naïve Bayes

4.1 Seleksi Atribut/Variabel

Terdapat 10 Variabel yang terdapat pada dataset yang akan diklasifikasikan ini, antara lain :

X1 = Type of School
X2 = School Accreditation
X3 = Gender
X4 = Interest
X5 = Residence
X6 = Parent Age
X7 = Parent Salary
X8 = House Area
X9 = Average Grades
X10 = Parent was in College

Sedangkan class dari dataset ini adalah :

C1 = in College : True ➔ 22 P(C1) = 22/50 = 0.44
C2 = in College : False ➔ 28 P(C2) = 28/50 = 0.56

4.2 Perhitungan Probabilitas Data Training

Perhitungan probabilitas P(Xk|Ci) untuk setiap class data training disajikan pada Tabel 3.

p-ISSN : 2087-0868 e-ISSN : 2598-9707

Tabel 3. Perhitungan P(Xk|Ci)

Atribut \ Class	Jumlah In College		Probabilitas in College (P(Xk|Ci)	
	True	False	True	False
Type of School = Academic	11	14	11/22 = 0.5	14/28 = 0.5
Type of School = Vocational	11	14	11/22 = 0.5	14/28 = 0.5
School Accreditation = A	15	11	15/22 = 0.681	11/28 = 0.392
School Accreditation = B	7	17	7/22 = 0.318	17/28 = 0.607
Gender = Male	11	14	11/22 = 0.5	14/28 = 0.5
Gender = Female	11	14	11/22 = 0.5	14/28 = 0.5
Interest = Very Interested	13	11	13/22 = 0.590	11/28 = 0.392
Interest = Less Interested	3	6	3/22 = 0.136	6/28 = 0.214
Interest = Quiet Interested	0	2	0	2/28 = 0.071
Interest = Not Interested	2	3	2/22 = 0.090	3/28 = 0.107
Interest = Uncertain	4	6	4/22 = 0. 181	6/28 = 0.214
Residence = Urban	14	10	14/22 = 0.636	10/28 = 0.357
Residence = Rural	8	18	8/22 = 0.363	18/28 = 0.642
Parent Age <= 50	5	11	5/22 = 0.227	11/28 = 0.392
Parent Age > 50	17	17	17/22 = 0.772	17/28 = 0.607
Parent Salary <= 5 Juta	3	15	3/22 = 0.136	15/28 = 0.535
Parent Salary >5Juta	19	13	19/22 = 0.863	13/28 = 0.464
House Area <=70	3	17	3/22 = 0.136	17/28 = 0.607
House Area >70	19	11	19/22 = 0.863	11/28 = 0.392
Average Grades <=86,5	9	21	9/22 = 0.409	21/28 = 0.75
Average Grades >86,5	13	7	13/22 = 0.590	7/28 = 0.25
Parent was in college = True	11	20	11/22 = 0.5	20/28 = 0.714
Parent was in college = false	11	8	11/22 = 0.5	8/28 = 0.285

4.3 Pengujian Data Testing

Data yang digunakan untuk data testing :

Tabel 4. Data Testing

Type school	school accre-ditta-tion	gender	interest	resi-dence	Parent age	Parent salary	House area	Average grades	Parent Was In college
Vocational	A	Male	Very Interested	Rural	<=50	>5,5 Juta	89,7	90,78	True
Vocational	A	Male	Less Interested	Urban	> 50	>5,5 Juta	61,6	87,56	False
Academic	B	Female	Less Interested	Urban	> 50	<= 5,5 Juta	57,2	84,37	False
Vocational	B	Male	Uncertain	Rural	> 50	>5,5 Juta	87,1	88,19	True
Vocational	A	Male	Very Interested	Urban	> 50	<= 5,5 Juta	88,2	93,05	False

p-ISSN : 2087-0868 e-ISSN : 2598-9707

Pengujian Data Testing

Tabel 5. Uji Data Testing 1

Class / Atribut	Probabilitas in College	
	True	**False**
Type of School = Vocational	0.5	0.5
School Accreditation = A	0.681	0.392
Gender = Male	0.5	0.5
Interest = Very Interested	0.590	0.392
Residence = Rural	0.363	0.642
Parent Age <= 50	0.227	0.392
Parent Salary >5Juta	0.863	0.464
House Area >70	0.863	0.392
Average Grades >86,5	0.590	0.25
Parent was in college = True	0.5	0.714
P(Xk|Ci)	$0.5 \times 0.681 \times 0.5 \times 0.590 \times 0.363 \times 0.227 \times 0.863 \times 0.863 \times 0.590 \times 0.5 = 0.00181$	$0.5 \times 0.392 \times 0.5 \times 0.392 \times 0.642 \times 0.392 \times 0.464 \times 0.392 \times 0.25 \times 0.714 = 0.00031$
P(X|Ci) x P(Ci)	$0.00181 \times 0.44 = 0.0007964$	$0.00031 \times 0.56 = 0.0001736$

Berdasarkan perhitungan yang telah dilakukan, maka Data Testing 1 memiliki class “in College = True” karena pada P(X|in College = “True”) memiliki nilai maksimum.

Tabel 6. Uji Data Testing 2

Class / Atribut	Probabilitas in College	
	True	**False**
Type of School = Vocational	0.5	0.5
School Accreditation = A	0.681	0.392
Gender = Male	0.5	0.5
Interest = Less Interested	0.136	0.214
Residence = Urban	0.636	0.357
Parent Age >50	0.772	0.607
Parent Salary >5Juta	0.863	0.464
House Area <=70	0.136	0.607
Average Grades >86,5	0.590	0.25
Parent was in college = False	0.5	0.285
P(Xk|Ci)	$0.5 \times 0.681 \times 0.5 \times 0.136 \times 0.636 \times 0.772 \times 0.863 \times 0.136 \times 0.590 \times 0.5 = 0.000394$	$0.5 \times 0.392 \times 0.5 \times 0.214 \times 0.357 \times 0.607 \times 0.464 \times 0.607 \times 0.25 \times 0.285 = 9{,}120\text{E}-05$
P(X|Ci) x P(Ci)	$0.000394 \times 0.44 = 0.000173$	$9{,}120\text{E}-05 \times 0.56 = 5{,}107\text{E}-03$

Berdasarkan perhitungan yang telah dilakukan, maka Data Testing 2 memiliki class “in College = True” karena pada P(X|in College = “True”) memiliki nilai maksimum.

p-ISSN : 2087-0868 e-ISSN : 2598-9707

Tabel 7. Uji Data Testing 3

Class / Atribut	Probabilitas in College	
	True	**False**
Type of School = Academic	0.5	0.5
School Accreditation = B	0.318	0.607
Gender = Female	0.5	0.5
Interest = Less Interested	0.136	0.214
Residence = Urban	0.636	0.357
Parent Age >50	0.772	0.607
Parent Salary <=5Juta	0.136	0.535
House Area <=70	0.136	0.607
Average Grades <=86,5	0.409	0.75
Parent was in college = False	0.5	0.285
P(Xk|Ci)	$0.5 \times 0.318 \times 0.5 \times 0.136 \times 0.636 \times 0.772 \times 0.136 \times 0.136 \times 0.409 \times 0.5 = 2{,}0079\text{E}-05$	$0.5 \times 0.607 \times 0.5 \times 0.214 \times 0.357 \times 0.607 \times 0.535 \times 0.607 \times 0.75 \times 0.285 = 0{,}000488$
P(X|Ci) x P(Ci)	$2{,}0079\text{E}-05$ x 0.44 $= 8{,}8350\text{E}-06$	0,000488 x 0.56 $= 0{,}000274$

Berdasarkan perhitungan yang telah dilakukan, maka Data Testing 3 memiliki class “in College = False” karena pada P(X|in College = “True”) memiliki nilai maksimum.

Tabel 8. Uji Data Testing 4

Class / Atribut	Probabilitas in College	
	True	**False**
Type of School = Vocational	0.5	0.5
School Accreditation = B	0.318	0.607
Gender = Male	0.5	0.5
Interest = Uncertain	0.181	0.214
Residence = Rural	0.363	0.642
Parent Age >50	0.772	0.607
Parent Salary >5Juta	0.863	0.464
House Area >70	0.863	0.392
Average Grades >86,5	0.590	0.25
Parent was in college = True	0.5	0.714
P(Xk|Ci)	$0.5 \times 0.318 \times 0.5 \times 0.181 \times 0.363 \times 0.772 \times 0.863 \times 0.863 \times 0.590 \times 0.5 = 0{,}000886$	$0.5 \times 0.607 \times 0.5 \times 0.214 \times 0.642 \times 0.607 \times 0.464 \times 0.392 \times 0.25 \times 0.714 = 0{,}000411$
P(X|Ci) x P(Ci)	0,000886 x 0.44 $= 0{,}00039$	0,000411 x 0.56 = 0,00023

Berdasarkan perhitungan yang telah dilakukan, maka Data Testing 4 memiliki class “in College = True” karena pada P(X|in College = “True”) memiliki nilai maksimum.

p-ISSN : 2087-0868 e-ISSN : 2598-9707

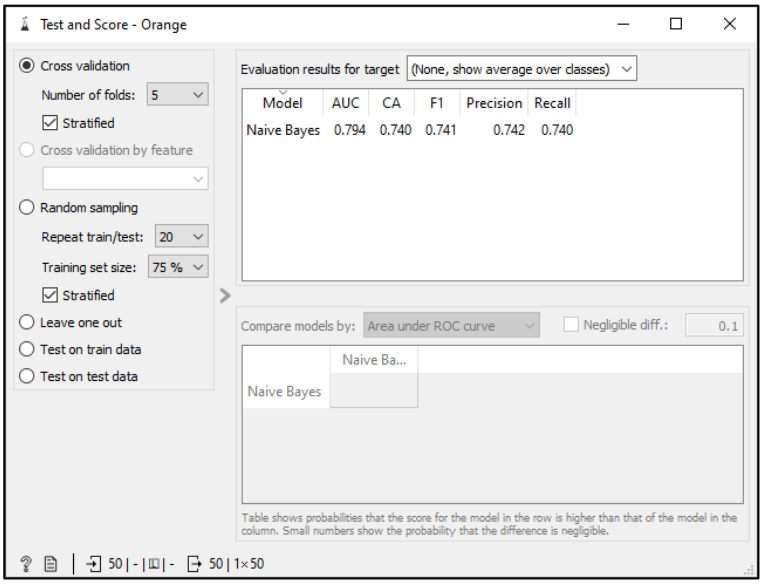
Tabel 9. Uji Data Testing 5

Class / Atribut	Probabilitas in College	
	True	**False**
Type of School = Vocational	0.5	0.5
School Accreditation = A	0.681	0.392
Gender = Male	0.5	0.5
Interest = Very Interested	0.590	0.392
Residence = Urban	0.636	0.357
Parent Age >50	0.772	0.607
Parent Salary <=5Juta	0.136	0.535
House Area >70	0.863	0.392
Average Grades >86,5	0.590	0.25
Parent was in college = False	0.5	0.285
P(Xk|Ci)	$0.5 \times 0.681 \times 0.5 \times 0.590 \times 0.636 \times 0.772 \times 0.136 \times 0.863 \times 0.590 \times 0.636 = 0{,}001708$	$0.5 \times 0.392 \times 0.5 \times 0.392 \times 0.357 \times 0.357 \times 0.607 \times 0.535 \times 0.392 \times 0.25 \times 0.285 = 0{,}0001244$
P(X|Ci) x P(Ci)	$0{,}001708$ x 0.44 $= 0{,}000751$	$0{,}0001244$ x 0.56 $= 6{,}966E-05$

Berdasarkan perhitungan yang telah dilakukan, maka Data Testing 5 memiliki class “in College = True” karena pada P(X|in College = “True”) memiliki nilai maksimum

4.4 Evakuasi Model

Evaluasi model dilakukan untuk mengetahui bagaimana performa model Naïve Bayes terhadap data yang diproses. Pada penelitian ini, digunakan confusion matrix dengan menggunakan aplikasi orange.



Gambar 3. Test & Score Akurasi Naïve Bayes

Berdasarkan pengolahan data menggunakan orange sebagaimana tersaji pada Gambar 3, model Naïve Bayes terhadap dataset memiliki akurasi sebesar 0.740 dengan AUC 0.794, F1 0.741, Precision 0.742, dan Recall 0.740. Hasil tersebut menunjukkan bahwa model Naïve Bayes cukup baik untuk mengklasifikasikan prediksi siswa masuk perguruan tinggi atau tidak (in college : true atau in college:false).

p-ISSN : 2087-0868 e-ISSN : 2598-9707

5. KESIMPULAN DAN SARAN

Proses data mining menggunakan Algoritma Naïve Bayes ini bertujuan untuk mempresiksi apakah seorang siswa akan melanjutkan ke perguruan tinggi atau tidak. Dataset yang digunakan sebagai data training berjumlah 50 record, sedangkan data testing 5 record. Untuk tingkat akurasi Naïve Bayes terhadap dataset berdasarkan perhitungan pada orange adalah sebesar 0,740. Untuk memperbesar tingkat akurasinya, diperlukan data training yang lebih banyak lagi.

6. DAFTAR PUSTAKA

[1] G. W. N. Wibowo, "Prediksi Kelanjutan Studi Siswa Ke Perguruan Tinggi Dengan Naïve Bayes," *J. DISPROTEK*, vol. 11, no. 1, pp. 41–46, 2020, doi: 10.34001/jdpt.v11i1.1159.

[2] M. F. Nugroho and S. Wibowo, "Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naïve Bayes," *J. Inform. Upgris*, vol. 3, no. 1, pp. 63–70, 2017, doi: 10.26877/jiu.v3i1.1669.

[3] E. Sutoyo and A. Almaarif, "Educational Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritme Naïve Bayes Classifier," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 95–101, 2020, doi: 10.29207/RESTI.V4I1.1502.

[4] A. Saleh, "Implementasi Metode Klafisikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 1, no. 2, pp. 73–81, 2019, doi: 10.20895/inista.v1i2.73.

[5] M. R. Handoko and Neneng, "Sistem Pakar Diagnosa Penyakit Ispa Menggunakan Metode Naïve Bayes Classifier Berbasis Web," *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 10, no. 3, p. 127, 2021, doi: 10.22303/csrid.10.3.2018.127-138.

[6] N. R. Indraswari and Y. I. Kurniawan, "Aplikasi Prediksi Usia Kelahiran Dengan Metode Naïve Bayes," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 9, no. 1, pp. 129–138, 2018, doi: 10.24176/simet.v9i1.1827.

NOMENKLATUR

X : Data dengan class yang belum diketahui
H : Hipotesis dan merupakan suatu class spesifik
$P(H|X)$: Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)
$P(H)$: Probabilitas hipotesis H (prior probabilitas)
$P(X|H)$: probabilitas X berdasarkan kondisi pada hipotesis H
$P(X)$: Probabilitas X