Jurnal ELTIKOM, Vol. 2, No. 2, Desember 2018, hal. 67-77
ISSN 2598-3245 (Print), ISSN 2598-3288 (Online)
Tersedia online di http://eltikom.poliban.ac.id
DOI : http://doi.org/10.31961/eltikom.v2i2.86

PERBANDINGAN ALGORITMA UNTUK KLASIFIKASI NILAI PADA PENJURUSAN SISWA SMA

Abdul Rahman Kadafi
STMIK Nusa Mandiri Jakarta
e-mail: abdurrahman.kadafi@gmail.com

ABSTRACT

The majors determination in senior high school usually using the recommendation of psicotest result, academic score, readability and talent of student. In this research, focused to compile the result of student academic score, for science and social knowledge. Which is the academic score that have the high influence in the majors of student at senior high school. There is some of algorithm that can help the classification process of data student for recommendation of majors, example C4.5, Naive Bayes,K-NN, Rule Induction, and each others. To know the validation level using cross validation. Then using T-Test to know the significans comparative of algorithm. The results of comparative analysis in this research, can be concluded that the Naïve Bayes algorithm method be the best algorithm than another algortihm, with the accuracy level on 79,51% and AUC value on 0,861.

Keywords*: Data Mining, Classification, Algorithm**, Majoring in Senior High School.*

ABSTRAK

Penentuan jurusan siswa pada tingkat pendidikan sekolah menengah atas pada umumnya menggunakan rekomendasi hasil psikotes, nilai akademik, minat dan bakat siswa. Tidak semua sekolah memiliki data yang lengakap untuk melakukan penjursan siswa. Dalam penelitian ini, difokuskan untuk mengomparasikan hasil nilai akademik siswa, untuk mata pelajaran rumpun ilmu pengetahuan alam dan ilmu pengetahaun sosial pada kelas 10 SMA. Nilai dari mata pelajaran yang manakah yang memiliki pengaruh tinggi terhadap penjurusan siswa di SMA. Terdapat beberapa algoritma dapat digunakan untuk membantu proses klasifikasi data siswa untuk rekomendasi penjurusan, misalnya C4.5, Naïve Bayes, K-NN, Rule Induction, dan lain-lain. Untuk mengetahui tingkat validasi digunakan metode cross validation. Kemudian digunakan T-Test untuk mengetahui signifikansi perbedaan antar algoritma. Hasil analisa komparasi pada penelitian komparasi algoritma untuk penjurusan ini, bahwa metode algoritma Naïve Bayes sebagai algoritma yang paling baik dibandingkan algoritma yang lainnya, yang meiliki akurasi pada 79,51% dan AUC pada nilai 0,861.

Kata Kunci*: Data Mining, Klasifikasi, Algoritma, Penjurusan SMA.*

I. PENDAHULUAN

PENGEMBANGAN sistem informasi pada lembaga pendidikan formal sudah menjadi kebutuhan yang selayaknya menjadi prioritas dari manajemen sekolah. Bantuan sistem informasi di sekolah, diharapkan dapat memeberikan kemudahan kepada manajemen sekolah dalam menjalankan manajerial pendidikan. Proses penerimaan siswa baru, proses pembelajaran, pengadaan bahan ajar dan administrasi diharapkan menjadi lebih mudah dengan menggunakan bantuan sistem informasi yang dirancang sesuai kebutuhan manajemen sekolah.

Salah satu proses yang dilaksanakan pada sekolah menengah atas (SMA) adalah penjurusan siswa. Proses penjurusan pada siswa menjadi hal yang tidak mudah jika sekolah tidak memiliki data siswa yang lengkap dan valid, yang dapat membantu penentuan penjurusan siswa. Penentuan penjurusan menjadi hal yang penting dikarenakan ini menjadi salah satu langkah awal penentuan jurusan kelak jika siswa melanjutkan ke perguruan tinggi. Penentuan jurusan siswa pada sekolah tingkat SMA dapat menggunakan rekomendasi hasil psikotes, nilai akademik, minat dan bakat siswa. Dalam penelitian ini, difokuskan untuk mengetahui mata pelajaran yang paling dominan terhadap penjurusan. Mata pelajaran yang menjadi atribut dalam penelitian tentang penjurusan siswa SMA ini adalah semua mata pelajaran rumpun IPA dan IPS pada kelas 10. Dari mata pelajaran yang digunakan untuk penentuan penjurusan

di SMA, terdapat mata pelajaran yang dominan dan berpengaruh siginifikan terhadap penentuan penjurusan siswa.

Dengan menerapkan teknik klasifikasi data mining, dapat digali informasi yang menjadi salah satu bahan rekomendasi penjurusan siswa SMA. Informasi ini dapat digunakan sebagai penunjang data yang dimiliki oleh pihak sekolah. Terdapat beberapa algoritma yang dapat membantu proses klasifikasi data siswa untuk rekomendasi penjurusan, misalnya *C4.5, Naïve Bayes, K-NN, Rule Induction,* dan lain-lain.

Algoritma *Decision Tree* menjadi lagoritma yang memiliki keakuratan paling bagus diantara algoritma yang lainnya [1]. Algoritma klasifikasi C4.5 dapat digunakan sebagai algoritma klasifikasi dalam data mining, yang dapat megnhasilkan pohon keputusan [2]. Dari beberapa algoritma klasifikasi, dapat digunakan sesuai kebutuhan klasifikasi data. Penelitian ini membandingkan akurasi hasil metode klasifikasi rekomendasi penjurusan siswa SMA, yaitu algoritma *C4.5, Naïve Bayes, K-NN,* dan *Rule Induction*.

II. LANDASAN TEORI

Masalah klasifikasi adalah mempelajari struktur kumpulan data contoh, yang sudah dipartisi menjadi beberapa kelompok, yang disebut kategori atau kelas (Aggarwal, 2015) [3]. Raghu Ramakrishnan (2004) Proses *Knowledge discovery* dan data mining (KDD) kira-kira dapat dipisahkan menjadi empat langkah[4]:

1. Seleksi Data: subset dan atribut interest target diidentifikasikan dengan memeriksa keseluruhan dataset yang belum diproses.
2. Pembersihan Data: Output data yang tidak relevan dan nilai yang tidak homogeny secara statistic dapat dibuang, nilai field dapat ditransformasikan ke unit umum dan beberapa field baru dibuat dengan mengkombinasikan field yang adauntuk memfasilitasi analisis.
3. Data Mining: mengaplikasikan algoritma data mining untuk mengekstrak pola yang menarik
4. Evaluasi: pola dipresentasikan kepada pengguna akhir dalam bentuk yang dapat difahami.

A. Teknik Klasifikasi

Klasifikasi data merupakan proses yang dilakukan untuk menemukan pola karakteristik himpunan data dalam sebuah basis data dan mengklasifikasikan ke dalam kelompok yang berbeda sesuai dengan model klasifikasi yang dihasilkan. Dalam proses klasifikasi digunakan data training dari atribut yang tersedia untuk menghasilkan model tertentu. Model yang dihasilkan kemudian akan digunakan untuk mengklasifikasikan data yang berbeda yang belum diketahui pola sebelumnya. Pohon keputusan merupakan salah satu teknik dalam klasifikasi data [2]. Dalam penelitian ini membandingkan implementasi algoritma klasifikasi menggunakan software RapidMiner sebagai berikut:

a. Naïve Bayes

Naïve Bayes merupakan metode dengan melakukan klasifikasi probabilistik sederhana dengan menghitung peluang berdasarkan jumlah frekuensi dan kombinasi dari data yang ada. Algoritma ini menggunakan prinsip teori *Bayes* dengan asumsi bahwa semua atribut tidak saling bergantung satu sama lain [7]. Persamaan yang digunakan dalam algoritma *Naïve Bayes* adalah sebagai berikut [7]:

$$P(H|X) = \frac{P(X|H) \, . P(H)}{P(X)} \quad (1)$$

Di mana:

X : Data dengan kelas yang belum diketahui
H : Hipotesa data kelas
P(H|X) : Probabilitas hipotesa H berdasar kondisi X
P(H) : Probabilitas hipotesa H
P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) : Probabilitas X

Dalam penerapan metode *Naive Bayes*, diperlukan beberapa petunjuk dalam proses klasifikasi untuk menentukan kelas yang sesuai dengan data sampel yang dianalisa. Metode *Naive Bayes* dapat dinyatakan dalam persamaan berikut [7]:

$$P(C|XF1 \dots Fn) = \frac{P(C)P(F1 \dots Fn|C)}{P(F1 \dots Fn)} \quad (2)$$

Dari persamaan diatas, variabel C menjelaskan kelas, sedangkan variabel F_1 ... F_n menggambarkan kondisi yang dibutuhkan untuk proses klasifikasi data. Rumus pada persamaan tersebut menjelaskan bahwa probabilitas masuknya sampel karakteristik tertentu ke dalam kelas C (*Posterior*) merupakan probabilitas munculnya kelas C, dikalikan dengan probabilitas munculnya karakteristik sampel pada keals C (*likelihood*), dibagi dengan probabilitas munculnya karakteristik sampel secara keseluruhan. Sehingga rumus pada persamaan 2 diatas, dapat dituliskan dalam bentuk yang lebih sederhana sebagai berikut [7]:

Posterior = (prior x likelihood) / evidence (3)

Nilai dari *Evidence* tetap pada setiap kelas untuk satu sampel data tertentu. Untuk menentukan ke kelas apa sampel tertntu akan diklasifikasikan, nilai pada *posterior* akan dibandingkan dengan nilai posterior pada kelas yang lainnya. Penjabaran dari rumus *Bayes* menggunakan aturan perkalian sebagai berikut ($C|F1,\dots,$):

$$\begin{aligned}(C|F1,\dots, &= P(C)P(F1,\dots,Fn|C)\\ &= (C)\ (F1|C)\ (F2,\dots,|C,F1)\\ &= (C)\ (F1|C)\ (F2|C,1)\ P(F3,\dots,Fn|C,F1,F2\\ &= (C)\ P(F1|C)P(F2|C,F1)\ P(F3|C,F1,F2)\ P(F4,\dots,Fn|C,F1,F2,F3)\\ &= P(C)\ P(F1|C)P(F2|C,F1)\ P(F3|C,F1,F2)\ \dots\ P(Fn|C,F1,F2,F3,\dots,Fn-1)\end{aligned} \quad (5)$$

Dari rumus yang dikemukakan pada persamaan 5 diatas, dapat dijelaskan bahwa faktor yang mempengaruhi nilai probabilitas semakin banyak dan beragam, dan tidak mudah dilakukan analisa satu persatu. Sehingga perhitungan probabilitas pada persamaan tersebut, sangat sulit dilakukan. Digunakan asumsi independen yang sangat tinggi (naif), masing-masing faktor *(F1,F2...Fn)* tidak salig terikat satu sama lain *(independen).* Berdasarkan asumsi tersebut, maka persamaan 5 dapat diberlakukan menjadi kesamaan berikut:

$$P(Fi|Fj) = \frac{P(Fi \cap Fj)}{P(Fj)} = \frac{P(Fi)\ P(Fj)}{P(Fj)} = P(Fi) \quad (6)$$

Dengan i≠j, dan
$P(Fi|C,Fj) = P(Fi|C)$

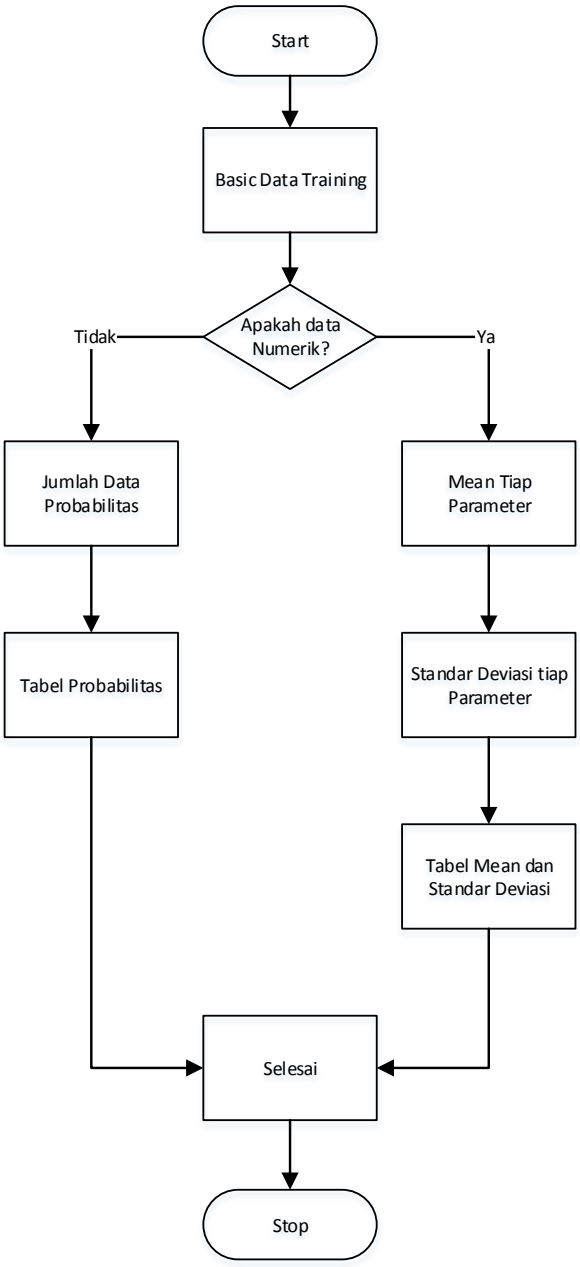
Persamaan 6 di atas adalah model teori *Naive Bayes* yang akan digunakan untuk proses klasifikasi dalam data mining. Sedangkan klasifikasi untuk data kontinyu dapat digunakan rumus *Densitas Gauss [6]*:

$$P(X_i = x_i|Y = y_j = \frac{1}{\sqrt{2\pi\sigma_{ij}}}\ e^{-\frac{(x_{i-}\mu_{ij})2}{2\sigma^2 ij}} \quad (7)$$

Keterangan:
P : Peluang
Xi: Atribut ke *i*
Xi: Nilai atribut ke *i*
Y : Kelas yang dicari

Yi: Sub kelas *Y* yang dicari
M: *mean*, menyatakan rata – rata dari seluruh atribut
σ : Standar deviasi , menyatakan varian dari seluruh atribut.

Dalam proses klasifikasi data, alur penerapan metode *Naive Bayes* dapat dijelaskan melalui bagan berikut.



Gambar 1. Alur Algoritma *Naïve Bayes* [6]

Dari Gambar 1 dapat dijelaskan bahwa alur kerja metode *Naïve Bayes* adalah sebagai berikut:
1. Mempersiapkan *data training*
2. Menghitung Jumlah dan probabilitas, jika data bertipe numerik maka:
   a. Mencari nilai *mean* dan standar *deviasi* dari parameter yang bertipe data *numeric*, dengan persamaan yang digunakan untuk menghitung nilai *mean* adalah sebagai berikut:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \quad (8)$$

Atau

$$\mu = \frac{(x1 + X2 + x3 \ldots + xn)}{n} \quad (9)$$

dengan:
μ : nilai rata-rata (*mean*)
xi : nilai *sample* ke -*i*
n : jumlah sampel yang digunakan

persamaan yang dapat digunakan untuk menghitung nilai standar deviasi adalah sebagai berikut:

$$\sigma = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \mu)^2}}{n-1} \quad (10)$$

Dengan keterangan:
σ : standar deviasi
x_i : nilai x ke -*i*
μ : nilai rata-rata
n : jumlah sampel yang digunakan

b. Menghasilkan nilai rata-rata, standar deviasi dan probabilitas.
3. jika data yang diolah tidak bertipe numerik maka langkah berikutnya adalah mencari nilai *probabilistic*, jumlah data sesuai dari kategori yang sama dihitung kemudian dibagi dengan jumlah data pada kategori tersebut.
4. Menghasilkan solusi.

b. K-Nearest Neighbor

Salah satu algoritma dalam data mining dengan jenis *supervised learning* adalah Algoritma *k-NN*. Seperti halnya algoritma klasifikasi lainnya, algoritma K-NN digunakanuntuk mengklasifikasikan data atau objek tertentu berdasarkan atribut pada data training yang menjadi sampel data, berdasarkan sebagian besar data dari kategori K-NN. Sesuai dengan namanya, algoritma K-NN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi terdekat dari uji sampel yang akan diprediksikan, dengan jarak yang digunakan adalah *Euclidean Distance*, yang dapat didefinisikan sebagai berikut [8]

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(a_r(x_i) - a_r(x_j))^2} \quad (11)$$

dengan :
d(xi,xj) : *Euclidean (Eucledean Distance).*
(xi) I : *record data* ke- *i*
(xj) : *record data* ke- *j*
a(r) : data ke-*r*
i,j :1,2,3,…n

Dalam algoritma *k*-NN penentuan nilai jarak saat dilakukan pengujian *data testing* dengan *data training* digunakan nilai terkecil dari nilai ketetanggaan terdekat [8] dari data tertentu yang dapat didefinisikan sebagi berikut:

$$D_{nn}(C_1, C_2) = min_{1 \le i \le r, 1 \le j \le s} d(y_i, z_j) \quad (12)$$

c. C4.5

Sebagai perbaruan dari algoritma pendahulunya ID3, algoritma C4.5 memiliki prinsip kerja yang hampir sama dengan algoritma ID3, yang menggunakan *Gain Ratio* sebagai indikator yang digunakan dalam pemilihan atribut pada proses klasifikasi. Rumus perhitungan gain ratio adalah sebagai berikut [9]:

$$Gain\ Ratio\ (S,A) = \frac{Gain\ (S,A)}{SplitInfo\ (S,A)} \quad (13)$$

Dari perhitungan yang dilakukan menggunakan rumus apda persamaan 13 diatas, maka atribut yang memiliki *Gain Ratio* tertinggi akan terpilih sebagai *root*. Selain *Gain Ratio*, pada C4.5 digunakan *SplitInfo* untuk menyatakan entropi dari masing-masing atribut, dengan persamaan yang digunakan adalah sebagai berikut [9]:

$$SplitInfo(S,A) = -\sum_{i=1}^{k} \frac{S_i}{S} log_2 \frac{S_i}{S} \quad (14)$$

Jika dibandingkan dengan ID3, C45 memiliki kelebihan yaitu dapat digunakan untuk mengolah data *numeric* dan diskret, dapat menangani data yang missing, dan menghasilkan rule-rule yang mudah untuk diinterpretasikan dalam tahapan *deployment* [11].

d. Rule Induction

Salah satu algoritma yang masuk dalam kelompok klasifikasi adalah Rumus *Rule Induction*, dengan ketentuan yang digunakan dalam proses klasifikasi adalah sebagai berikut [12]:

a. Untuk mendapatkan nilai support untuk sebuah item A
   Support = jumlah transaksi yang mengandung item A / Total transaksi
b. Untuk mencari nilai support dari 2-item
   Support (A, B) = P (A ∩ B)
   P (A ∩ B) = Jumlah transaksi yang mengandung A dan B / Total Transaksi
c. Mencari nilai *confidance*
   Confidance (A → B) = P (A | B)
   P (A | B) = Jumlah transaksi yang mengandung A dan B / Jumlah transaksi yang mengandung item A. Peoses perhitungan pada algoritma Rule Induction, menggunakan *information gain*.

B. Performansi

Untuk mengevaluasi model yang dihasilkan oleh algoritma klasifikasi yang digunakan, digunakan *F-Mearsure*. *F-measure* dapat digunakan untuk melakukan evaluasi dalam menemukan informasi yang dikombinaksikan melalui *recall* dan *precision*. Dari hasil perhitungan yang didapatkan, nilai *recall* dan *precision* dapat memiliki bobot tidak selalu sama. Berikut adala cara menentukan akurasi dengan menggunakan *F-Measure* [14].

TABEL I
F-MEASURE

		Aktual	
		Tidak *Churn*	*Churn*
Prediksi	Churn	a	b
	Tidak Churn	c	d

1. $TP\ (True\ Positif) = \frac{a}{(c+d)}\ x\ 100\%$
2. $FP\ (False\ Positif) = \frac{b}{(a+b)}\ x\ 100\%$
3. $TN\ (True\ Negatif) = \frac{a}{(a+b)}\ x\ 100\%$
4. $FP\ (False\ Positif) = \frac{d}{(c+d)}\ x\ 100\%$
5. $Akurasi = \frac{(a+d)}{(a+b+c+d)}\ x\ 100\%$
6. $Precissision = \frac{TP}{(TN+FP)}\ x\ 100\%$
7. $Recall = \frac{TP}{(TP+FN)}\ x\ 100\%$

III. METODE PENELITIAN

A. Data dan Variabel

Sampel dari penelitian ini adalah data nilai siswa SMA kelas 10 sebanyak 131 data, dan data penjurusan siswa di kelas 11 SMA. Atribut dalam penelitian ini adalah nilai mata pelajaran rumpun IPA (Matematika, Fisika, Kimia dan Biologi) dan IPS (Sejarah, Geografi, Sosiologi, dan Ekonomi). Dengan label data penelitian adalah penjurusan siswa SMA pad akelas 11.

B. Model Validasi

Model validasi yang digunakan dalam penelitian ini adalah *cross validation* sebagai teknik yang digunakan untuk melakukan validasi terhadap proses klasifikasi yang menggunakan algoritma C4.5, K-NN, *Naïve Bayes*, dan *Rule Induction*. *Cross-validation* merupakan metode algoritma statistik dengan membagi data menjadi dua segmen yaitu data latih dan data tes untuk memvalidasi model yang diusulkan, yang menggunakan pengujian standar untuk mengetahui *error rate* [10]. Setiap data akan diproses secara acak dalam proporsi yang tepat untuk menjadi data training dan data testing, dengan perbandingan yang sama. Dalam proses validasi menggunakan cross validation, pengujian diulangi beberapa kali dengan menggunakan sample yang berbeda. Hasil dari proses iterasi dari masign-masing uji data, akan dirata-ratakan sehingga menghasilkan error rate secara keseluruhan (tabel 2) [14].

TABEL II
TABEL CROSS VALIDATION

Eksperimen	Dataset
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

C. Evaluasi Area Under Curve

Hasil dari tahap pengujian cross validation adalah nilai accuracy dari *confusion matrix* dan *area under curve* (AUC) dari ROC, yang digunakan sebagai indikator pengukuran tingkat akurasi performa dari algoritma yang diujikan. Nilai *accuracy* dalam penelitian ini diperoleh dari tabel *confusion matrix* dari hasil pengolahan menggunakan RapidMiner. AUC menunjukkan pengukuran performansi metode klasifikasi berdasarkan ROC *curve*. Skala dalam kualifikasi performansi AUC adalah 0 sampai 1, dimana angka 0 menunjukkan tingkat negatif dan angka 1 menunjukkan tingkat positif [15]. Berikut adalah tabel performansi dari AUC.

TABEL III
KLASIFIKASI PERFORMANCE AUC

Performance	Klasifikasi
0,90 – 1,00	Paling Baik
0,80 – 0,90	Baik
0,70 – 0,80	Adil / Sama
0,60 – 0,70	Rendah
0,50 – 0,60	Gagal

D. T-Test

Untuk mengetahui hasil perbedaan dari performasi algoritma yang dibandingkan digunakan *T-Test* untuk melakukan pengujian hipotesis. *T-Test* satu sampel tergolong hipotesis deskriptif, dengan rumus yang digunakan untuk perhitungannya adalah sebagai berikut [16].

$$T_{hitung} = \frac{X - \mu_0}{\frac{\sigma}{sn}} \quad (17)$$

T_{hitung}= Nilai T yang yang dicari dan menunjukan standar deviasi pada distribusi normal (tabel t)
X = Nilai rata-rata dari data yang diolah
μ_0 = Rata-rata nilai yang menjadi sampel
s = Standar deviasi dari populasi yang telah diketahui
n = Jumlah populasi penelitian
Jika nilai uji T adalah lebih kecil dari 0,05 maka Ho ditolak.

IV. HASIL DAN PEMBAHASAN

A. *Data Processing*

Dalam proses pengolahan data nilai siswa kelas 10 SMA yang digunakan sebagai sampel penelitian yang berjumlah 133 data nilai siswa, didapatkan statistik data sebagai berikut:

Name	Type	Missing	Statistics		Filter (17 / 17
Label JUR	Binominal	0	Least IPS (65)	Most IPA (66)	Values IPA (66), IPS (65)
MTK 1	Integer	0	Min 62	Max 96	Average 81.344
FIS 1	Integer	0	Min 53	Max 90	Average 79.664
BIO 1	Integer	0	Min 61	Max 91	Average 79.634
KIM 1	Integer	0	Min 64	Max 94	Average 81.794
SEJ 1	Integer	0	Min 54	Max 96	Average 84.366
GEO 1	Integer	0	Min 45	Max 95	Average 84.443
EKO 1	Integer	0	Min 44	Max 97	Average 82.534
SOS 1	Integer	0	Min 81	Max 98	Average 91.435
MTK 2	Integer	0	Min 76	Max 95	Average 81.954
FIS 2	Integer	0	Min 77	Max 92	Average 82.282
KIM 2	Integer	0	Min 77	Max 94	Average 80.954
BIO 2	Integer	0	Min 77	Max 92	Average 82.550
SEJ 2	Integer	0	Min 79	Max 92	Average 82.649
GEO 2	Integer	0	Min 78	Max 96	Average 84.366
EKO 2	Integer	0	Min 72	Max 95	Average 83.931
SOS 2	Integer	0	Min 82	Max 96	Average 88.718

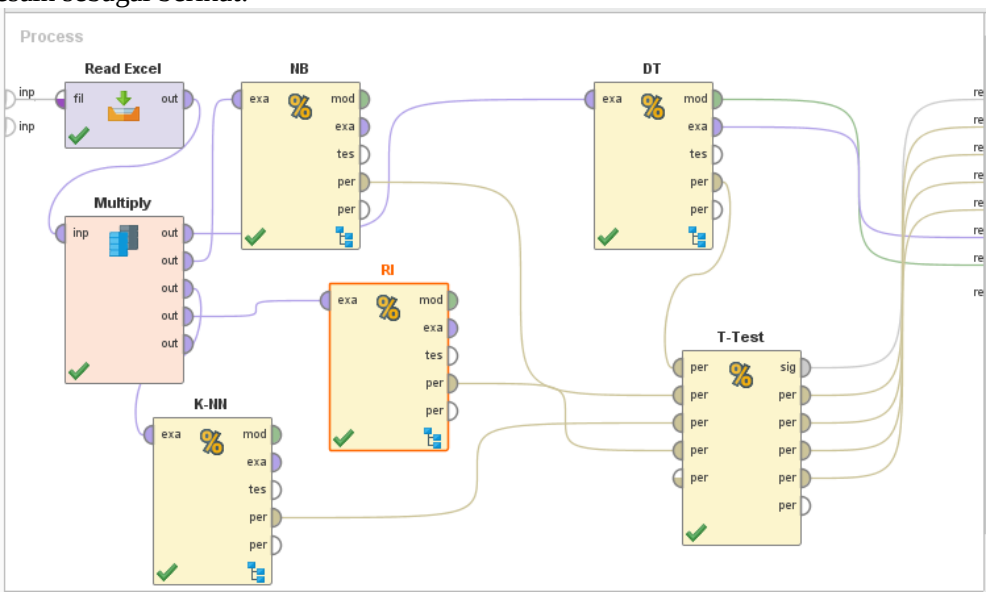
Gambar 2. Statistik data penelitian

ISSN 2598-3245 (Print)
ISSN 2598-3288 (Online)

Dari tabel yang tertera pada Gambar 2 di atas, dapat terlihat bahwa tidak ada data yang missing dari sampel yang digunakan, utnuk nilai siswa dari mata pelajaran rumpun IPA dan rumpun IPS pada kelas 10 SMA semester 1 dan 2.

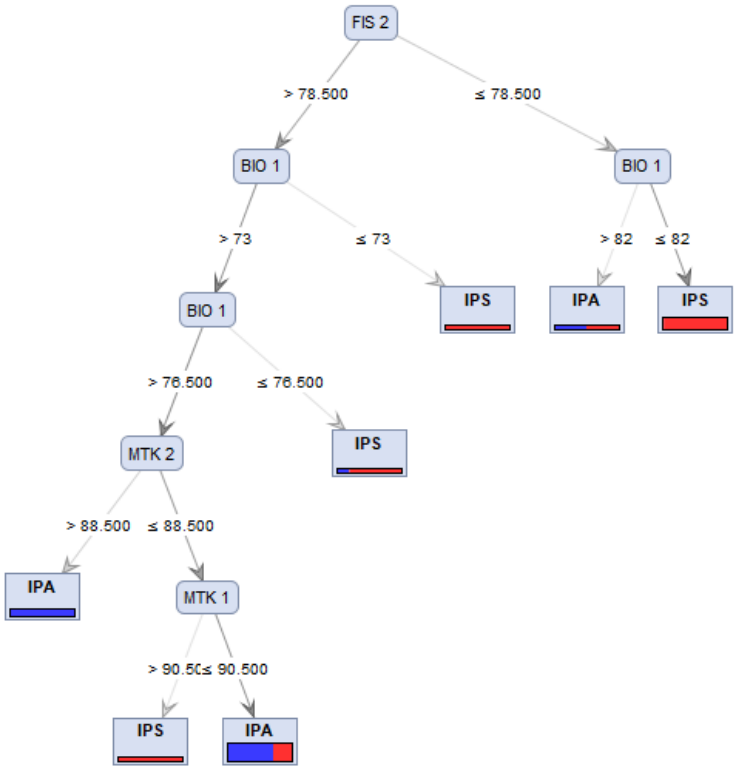
B. Modelling

Dalam proses pengolahan data menggunakan Rapidminer, digunakan empat algoritma klasifikasi untuk membandingkan keandalan dalam mengolah data sampel. Pemrosesan data dapat digambarkan dengan desain sebagai berikut:



Gambar 3. Desain pemrosesan data

Dari gambar 3 di atas terlihat bahwa dalam pemrosesan data yang digunakan dalam sampel penelitian, digunakan empat empat algoritma klasifikasi yaitu *Naïve Bayes*, C4.5, K-NN, dan *Rule Induction*. Dari pemrosesan berdasarkan desain pada gambar 3, didapatkan pola klasifikasi sebagai berikut:



Gambar 4. Pola klasifikasi penjurusan berdasarkan mata pelajaran

Keterangan:
Fis 2 : Fisika semester 2
Bio 1 : Biologi semester 1
MTK 1 : Matematika semester 1
MTK 2 : Matematika semester 2

Dari model decision tree diatas, dapat dijelaskan bahwa mata pelajaran yang paling berpengaruh dalam penentuan penjurusan siswa SMA adalah mata pelajaran Fisika. Untuk mata pelajaran yang dominan berikutnya adalah mata pelajaran Biologi, dan mata pelajaran yang dominan ketiga adalah Matematika. Mata pelajaran Kimia, tidak terlihat dominan dalam penentuan jurusan siswa SMA. Untuk mata pelajaran rumpun ilmu pengetahuan sosial, tidak terlihat memiliki pengaruh terhadap pemilihan penjurusan siswa SMA.

C. Evaluation

Dari hasil pengolahan dan pengujian menggunakan teknik Cross Validation, didapatkan hasil pengujian sebagai berikut:

TABEL IV
HASIL AKURASI DAN AUC

Aspek	C4.5	NB	K-NN	RI
Accuracy	77,09%	79,51%	77,86%	75,57%
AUC	0,768	0,861	0,500	0,790

Berdasarkan data diatas, terlihat bahwa *Naïve Bayes* memiliki nilai akurasi paling tinggi yakni 79,51%, kemudian diikuti oleh K-NN pada nilai 79,51%. Diposisi ketiga dari 4 algoritma yang dikomparasi adalah *C4.5* dengan nilai akurasi 77,09%, algoritma dengan akurasi paling rendah adalah Rule Induction dengan nilai akurasi 75,57%.

Sedangkan untuk evaluasi hasil implementasi algoritma, digunakan nilai *area under curve*. Nilai akurasi paling tinggi ada pada algoritma *Naïve Bayes* yakni 0,861 dengan kategori sangat baik, kemudian diikuti oleh algoritma *Rule Induction* dengan nilai *AUC* 0,790 dengan kategori baik. Pada urutan ketiga adalah algoritma *C4.5* dengan nilai *AUC* 0,768 dengan kategori netral. Sedangkan unutk algoritma K-NN memiliki nilai AUC 0,500 sehingga algoritma K-NN dinyatakan gagal.

Untuk mengetahui lebih lanjut perbandingan dari 5 algoritma tersebut, dilakukan *T-test*, dengan hasil sebagai berikut:

TABEL V
HASIL T-TEST

	C4.5	NB	K-NN	RI
C4.5		0,644	0,848	0,761
NB			0,694	0,417
K-NN				0,539
RI				

Berdasarkan tabel *T-test* diatas dapat dijelaskan bahwa hipotesa yang dihasilkan dari pengujian komparasi algoritma diatas ditolak. Meskipun dalam pengujian akurasi, *Naïve Bayes* memiliki akurasi dan nilai AUC yang paling tinggi artinya dari keempat algortima tersebut memiliki akurasi yang hampir sama, dan tidak memiliki perbedaan yang signifikan.

V. KESIMPULAN

Berdasarkan pengolahan data dan analisa yang dilakukan, dapat disimpulkan bahwa mata pelajaran yang paling dominan dalam penentuan penjurusan di SMA adalah mata pelajaran Fisika. Algoritma *Naïve Bayes* sebagai algoritma yang paling baik dibandingkan algoritma yang lainnya, dengan tingkat akurasi pada 79.51% dan AUC pada nilai 0,861.

DAFTAR PUSTAKA

[1] Nugroho, S. Yusuf, 2015, Klasifikasi dan Clustering Penjurusan Siswa SMA Negeri 3 Boyolali, Vol. 1 No. 1, Jurnal Ilmu Komputer dan Informatika, Universitas Muhammadiyah Surakarta.
[2] Dhika, Harry, 2015, Kajian Komparasi Penerapan Algoritma C4.5, Naïve Bayes, Neural Network Dalam pemilihan Mitra Kerja Penyedia Jasa Transportasi: Studi Kasus CV. Viradi Global Pratama, Prosiding Seminar Nasional Inovasi dan Tren (SNIT) 2015, Hal. A.
[3] Aggarwal, Charru, 2015, Data Mining: The Textbook, Ibm T.J. Watson Research Center Yorktown Heights New York Usa
[4] Ramakrishnan, Ragu; Johannes Gehrke, 2004, Sistem Manajemen Database, Edisi 3, Yogyakarta, Andi.
[5] Saleh, Alfa, 2015, Implementasi Klasifikasi Metode Naïve Bayes dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga, Citec Journal Vol. 2, No. 3 Mei 2015 – Juli 2015, ISSN : 2354 - 5771.
[6] Saleh , Alfa, 2015, Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga, Citec Journal, Vol. 2, No. 3, Mei 2015 – Juli 2015 ISSN: 2354-5771
[7] Pattekari, S. A., Parveen, A., 2012, Prediction System for Heart Disease Using Naive Bayes, *International Journal of Advanced Computer and Mathematical Sciences*, ISSN 2230-9624, Vol. 3, No 3, Hal 290-294.
[8] Krisandi, Nobertus, Dkk, 2013, Algoritma *K-Nearest Neighbor* Dalam Klasifikasi Data Hasil Produksi Kelapa Sawit Pada Pt. Minamas Kecamatan Parindu, Buletin Ilmiah Math. Stat. Dan Terapannya (Bimaster) Volume 02, No.1 (2013), Hal. 33-38.
[9] Agustina, Dana Melina, 2016, Analisis Perbandingan Algoritma ID3 Dan C4.5 Untuk Klasifikasi Penerima Hibah Pemasangan Air Minum Pada PDAM Kabupaten Kendal, Journal of Applied Intelligent System, Vol.1, No. 3, Oktober 2016 : 234-244
[10] Khasanah, Nidaul Fata, 2016, Klasifikasi Proses Penjurusan SiswaTingkat SMA Menggunakan Data Mining, Informatics For Educators And Professionals, Vol.1, No. 1, Desember 2016, 65 –69 E-Issn: 2548-3412,
[11] Selvia Lorena Br Ginting, 2014, Analisis Dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Memprediksi Masa Studi Mahasiswa Berdasarkan Data Nilai Akademik, Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST), ISSN: 1979-911
[12] Wulansari, Andhita Dessy, 2012, Perbandingan Ketepatan Klasifikasi Antara Metode Regresi Logistik Dan Klasifikasi Pohon Pada Kasus Program Wajardikdas 9 Tahun, Cendikia Vol. 10. No. 1, Jurusan Tarbiyah Stain Ponorogo
[13] Abdillah, Fakhrie, 2016, Penggunaan *Deep Learning* Untuk Prediksi *Churn* Pada Jaringan Telekomunikasi *Mobile*, e-Proceeding of Engineering : Vol.3, No.2 Agustus 2016, ISSN : 2355-9365, Bandung, Universitas Telkom.
[14] Hastuti, Khafiizh, 2012, Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif, Semarang, Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2012 (Semantik 2012) Isbn 979 - 26 - 0255 - 0
[15] Bustami, 2013, Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, *TECHSI: Jurnal Penelitian Teknik Informatika*, Vol. 3, No.2, Hal. 127-146.
[16] Nazir, M, 2007, "Metode Penelitian, Cetakan Ke Tiga", Jakarta, Ghalia Indonesia