

# Klasifikasi dan Klustering Penjurusan Siswa SMA Negeri 3 Boyolali

Yusuf S. Nugroho<sup>1\*</sup>, Syarifah N. Haryati<sup>1</sup>

<sup>1</sup>Program studi Informatika  
Universitas Muhammadiyah Surakarta  
Surakarta

\*Yusuf.Nugroho@ums.ac.id

## ABSTRAK

SMA N 3 Boyolali merupakan salah satu sekolah menengah di kota Boyolali yang saat ini telah memiliki 2 jurusan yaitu IPA dan IPS. Penjurusan siswa ini dapat mengarahkan peserta didik agar lebih fokus dalam mengembangkan kemampuan diri dan minat yang dimiliki. Pemilihan jurusan yang tidak tepat bisa sangat merugikan siswa terhadap minat dan karir mereka di masa mendatang. Dengan penjurusan tersebut diharapkan dapat memaksimalkan potensi, bakat atau talenta individu, sehingga dapat memaksimalkan nilai akademisnya. Berdasarkan latar belakang tersebut, maka dengan menerapkan teknik data mining diharapkan dapat membantu siswa untuk menentukan jurusan yang tepat sesuai dengan kriteria yang ditetapkan. Adapun teknik data mining yang digunakan dalam penentuan jurusan ini menggunakan 3 buah metode yaitu Algoritma C4.5, Naive Bayes dan Algoritma K-Means. Sedangkan atribut yang digunakan terdiri dari Gender, Minat, Rata-rata nilai IPA, Rata-rata nilai IPS, nilai Psikotest IPA, nilai Psikotest IPS, Asal Sekolah dan Jurusan. Analisis dilakukan dengan bantuan aplikasi RapidMiner 5 untuk mengetahui nilai-nilai perbandingan terhadap metode yang digunakan. Hasil penelitian menggunakan perbandingan 3 metode menunjukkan bahwa berdasarkan nilai precision, metode naive bayes lebih baik dibandingkan dengan metode yang lain dengan nilai 77,51%. Sedangkan berdasarkan nilai recall dan accuracy, decision tree lebih baik dibandingkan dengan metode yang lain dengan nilai recall sebesar 90,80% dan nilai accuracy sebesar 79,14%. Variabel yang paling berpengaruh dalam menentukan penjurusan yaitu rata-rata nilai IPA sehingga perlu dijadikan pertimbangan bagi pihak sekolah untuk menentukan jurusan siswa.

**Kata kunci:** algoritma C4.5, algoritma K-Means, data mining, jurusan SMA, naive bayes

## 1. PENDAHULUAN

Kemajuan teknologi informasi telah menyebabkan banyak orang dapat memperoleh data dengan mudah bahkan cenderung berlebihan. Data tersebut semakin lama semakin banyak dan terakumulasi, akibatnya pemanfaatan data yang terakumulasi tersebut menjadi tidak optimal [1]. Sebagai contoh perusahaan retail yang akan memberikan brosur penawaran barang-barang yang dijual ke pelanggan sesuai basis data pelanggan yang mereka punya. Jika perusahaan retail tersebut mempunyai satu juta data pelanggan dan masing-masing pelanggan tersebut dikirimkan sebuah brosur penawaran dimana biaya pengiriman brosur tersebut adalah dua ribu rupiah, maka biaya yang akan dikeluarkan oleh perusahaan tersebut adalah dua juta rupiah per bulan. Dari penggunaan dana tersebut mungkin hanya sepertiganya atau bahkan 8% saja yang secara efektif membeli penawaran tersebut [2].

SMAN 3 Boyolali yang berdiri sejak tahun 1989 merupakan salah satu Sekolah Menengah Atas di Kota Boyolali yang terletak di Jalan Perintis Kemerdekaan, Pulisen, Boyolali. Saat ini SMA tersebut telah memiliki

dua jurusan yaitu IPA dan IPS. Penjurusan siswa ini bertujuan untuk mengarahkan peserta didik agar lebih fokus mengembangkan kemampuan diri dan minat yang dimiliki. Jurusan yang tidak tepat bisa sangat merugikan siswa dan karirnya di masa mendatang. Dengan penjurusan tersebut diharapkan dapat memaksimalkan potensi, bakat atau talenta individu, sehingga juga akan memaksimalkan nilai akademisnya. Penentuan jurusan ini akan berdampak terhadap kegiatan akademik selanjutnya dan mempengaruhi pemilihan bidang ilmu atau studi bagi siswa-siswi yang ingin melanjutkan ke perguruan tinggi nantinya. Penentuan jurusan yang dilakukan selama ini mempunyai banyak kelemahan, antara lain berdasarkan keinginan siswa tanpa melihat latar belakang nilai akademisnya saja. Sehingga jurusan yang dipilih terkadang menjadi masalah bagi siswa di kemudian hari, sebagai contoh nilai akademik yang tidak maksimal, pemilihan program studi saat melanjutkan ke jenjang perguruan tinggi yang terkendala akibat jurusan SMA yang tidak sesuai, dan lain-lain.

Dengan menerapkan teknik klasifikasi dan klustering dalam *data mining* (DM), dapat digali suatu informasi

strategis yang dapat digunakan untuk menentukan penjurusan siswa. Informasi hasil analisis dapat digunakan untuk menemukan peluang-peluang yang baru serta menemukan rencana strategis dalam proses pengklasifikasian jurusan terhadap siswa.

Namun, kegiatan klasifikasi dan klustering jika dilakukan oleh manusia masih memiliki keterbatasan, terutama pada kemampuan manusia dalam menampung jumlah data yang ingin diolah. Selain itu bisa juga terjadi kesalahan akibat ketidakteelitian yang dilakukan. Salah satu cara mengatasi masalah ini adalah dengan menggunakan teknik *data mining* yang bisa digunakan untuk pengolahan data menjadi sumber informasi strategis dengan metode klasifikasi dan klustering. Data mining dapat membantu sebuah organisasi yang memiliki data melimpah untuk memberikan informasi yang dapat mendukung pengambilan keputusan [3].

Dalam bidang analisis perusahaan dan manajemen resiko, data mining digunakan untuk merencanakan keuangan dan evaluasi aset, merencanakan sumber daya (*resources planning*) dan memonitor persaingan [4].

Berdasarkan permasalahan tersebut, maka dalam penelitian ini dilakukan perbandingan 3 metode untuk penjurusan siswa menggunakan teknik *data mining*, yaitu metode *decision tree* dengan algoritma C.4.5, *naive bayes* dan *clustering* dengan algoritma *k-means*. Hasil yang diperoleh dapat memberikan informasi strategis bagi siswa yang dapat mendukung keputusan untuk menentukan jurusan serta memudahkan bagi pihak sekolah untuk mengarahkan siswa dalam hal penjurusan tersebut.

## 2. METODE

### 2.1 PENENTUAN ATRIBUTE

Atribut-atribut yang digunakan untuk perbandingan metode dalam *data mining* ini ditentukan sesuai dengan kebutuhan analisis. Adapun daftar atribut yang digunakan dapat dilihat pada tabel 1. Ada dua jenis variabel dalam proses ini, yaitu:

- a) Variabel dependen (Y)

Variabel dependen (Y) adalah variabel yang nilainya tergantung atau terikat berdasarkan nilai-nilai variabel lainnya.

- b) Variabel independen (X)

Variabel independen (X) adalah variabel yang nilainya tidak tergantung dari nilai-nilai variabel lainnya.

Tabel 1. Keterangan atribut yang digunakan

Atribut	Jenis Variable
Jurusan	Y
Gender	X1
Minat	X2
Rata-rata IPA	X3
Rata-rata IPS	X4
Psikotes IPA	X5
Psikotes IPS	X6
Asal sekolah	X7

### 2.2 PENENTUAN JUMLAH SAMPLE

Untuk mendapatkan sampel yang dapat menggambarkan dan mewakili jumlah populasi menggunakan bantuan metode slovin dengan nilai maksimal  $e = 5\%$ . Metode slovin dalam [5] ditunjukkan pada persamaan 1.

$$n = \frac{N}{1 + ne^2} \quad (1)$$

Keterangan:

n = jumlah sampel

N = jumlah keseluruhan data / populasi

e = galat kesalahan (ditentukan sebesar 5%)

### 2.3 PENGUMPULAN DATA

Jumlah data siswa SMAN 3 Boyolali selama 5 tahun diketahui sebanyak 1240 siswa. Dengan menggunakan persamaan 1, maka dapat dihitung jumlah sampel yang diambil yaitu sebanyak 302 siswa yang digunakan data *sampling*.

### 2.4 ANALISIS DATA

Tahap perbandingan metode *data mining* dilakukan dengan melakukan perhitungan menggunakan 3 algoritma yaitu algoritma C4.5, *Naive Bayes*, dan *K-Means*. Algoritma C4.5 merupakan salah satu algoritma dalam metode *decision tree* yang dihitung dengan penentuan atributnya menggunakan *information gain* berdasarkan entropi dari masing-masing atribut yang telah ditentukan dengan persamaan 2 dan 3 [6].

$$Entropy(y) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \dots - p_n \log_2 p_n \quad (2)$$

$$gain(y, A) = entropy(y) - \sum_{c \in nilai(A)} \frac{y_c}{y} entropy(y_c) \quad (3)$$

Sementara itu, algoritma *naive bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik untuk memprediksi peluang di masa depan [7]. Adapun algoritma ini dapat dihitung menggunakan persamaan 4.

$$P(S|X) = \frac{P(X|H).P(H)}{P(X)} \quad (4)$$

Sedangkan algoritma *K-Means* merupakan salah satu metode *clustering* non-hirarki yang mengelompokkan data

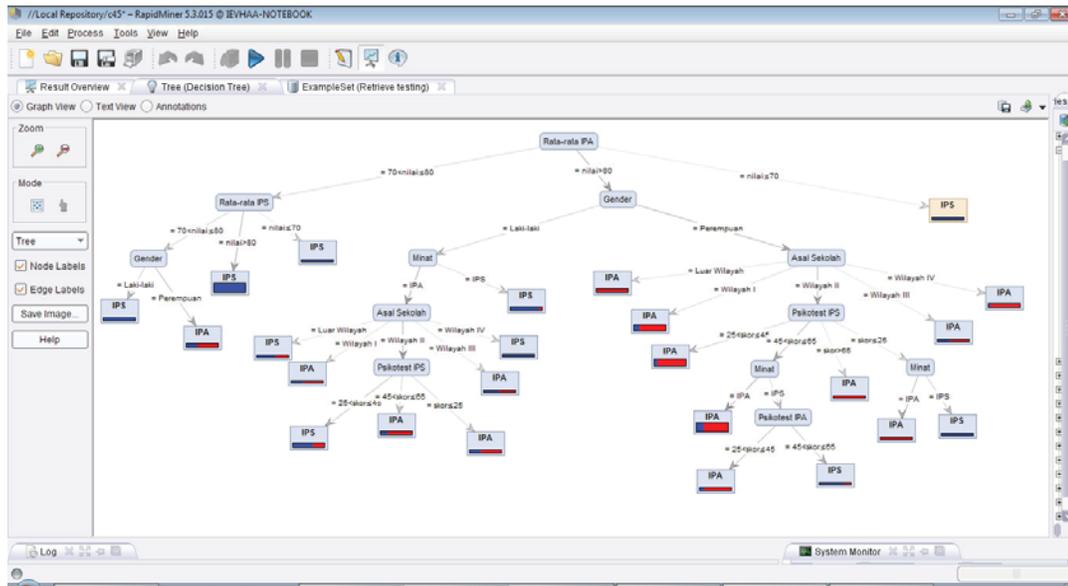
dalam bentuk satu atau lebih kluster [8]. Menurut [1], metode *K-Means* dapat mempartisi data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam cluster yang lain.

Pembagian kelompok kluster menggunakan persamaan *Euclidean distance space* (persamaan 5) yang sering digunakan dalam perhitungan jarak, hal ini dikarenakan hasil yang diperoleh merupakan jarak terpendek antara dua titik yang diperhitungkan.

$$du = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} \tag{5}$$

### 3.2 DISTRIBUSI PROBABILITAS PENJURUSAN SISWA

Distribusi data klasifikasi penjurusan siswa menggunakan *naive bayes* berdasarkan nilai rata-rata IPA dan IPS ditunjukkan pada gambar 2. Gambar tersebut menunjukkan bahwa distribusi terbanyak untuk jurusan IPS berasal dari siswa yang memiliki nilai rata-rata IPA  $70 < \text{nilai} \leq 80$  dan nilai rata-rata IPS  $70 < \text{nilai} \leq 80$ .

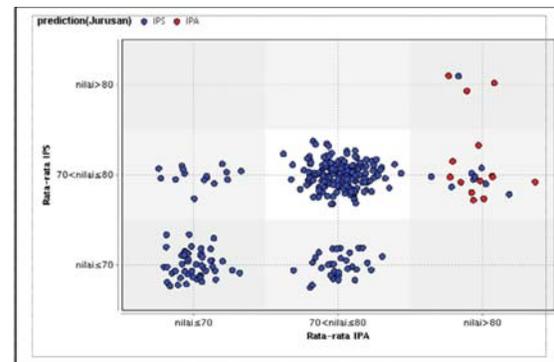


Gambar 1. Pohon keputusan untuk klasifikasi penjurusan siswa

## 3. HASIL

### 3.1 POHON KEPUTUSAN PENJURUSAN SISWA

Hasil proses klasifikasi penjurusan siswa dengan algoritma C4.5 ditunjukkan pada gambar 1. Berdasarkan hasil pohon keputusan pada gambar 1, dapat dilihat bahwa atribut yang memiliki pengaruh paling tinggi untuk menentukan klasifikasi penjurusan siswa adalah nilai rata-rata IPA. Hal ini ditunjukkan dengan variabel tersebut yang menempati sebagai simpul akar (*root node*).

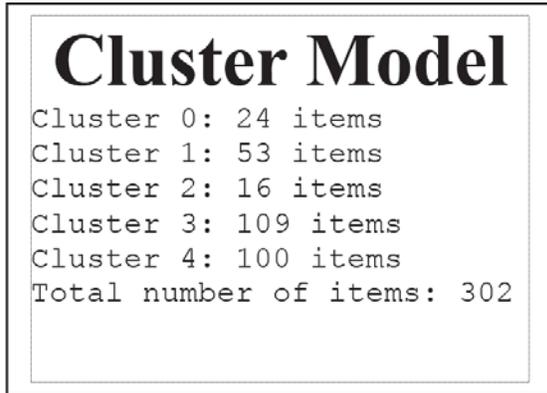


Gambar 2. Probabilitas penjurusan siswa menggunakan Naive Bayes

### 3.3 KLASTERING PENJURUSAN SISWA

Hasil klustering menunjukkan bahwa kelompok penjurusan siswa berjumlah 5 kluster dengan masing-masing jumlah data berbeda antara satu kluster dengan kluster yang lainnya seperti ditunjukkan pada gambar 3. Berdasarkan pembagian kluster tersebut dapat dilihat

bahwa kluster yang memiliki anggota terbanyak adalah kluster 3 yaitu sebanyak 109 dari 302 data, sedangkan yang memiliki anggota paling sedikit adalah kluster 2 yaitu sebanyak 16 dari 302 data.



Gambar 3. Cluster Model Text View menggunakan K-Means

#### 4. DISKUSI

##### 4.1 PERHITUNGAN ALGORITMA C4.5

Berdasarkan hasil pohon keputusan pada gambar 1 dapat diketahui bahwa rata-rata IPA memiliki pengaruh paling tinggi untuk menentukan jurusan di SMA tersebut. Hal ini ditunjukkan dengan atribut rata-rata IPA menempati sebagai *root node*. Adapun rata-rata IPS sebagai *internal root* pertama pada rata-rata IPA  $70 < \text{nilai} \leq 80$ , gender sebagai internal root pertama pada rata-rata IPA  $\text{nilai} > 80$  sedangkan pada rata-rata IPA  $\text{nilai} \leq 70$  tidak terdapat *internal node* dikarenakan hasil entropi dari variabel tersebut adalah 0, sehingga menghasilkan *leaf node* yaitu IPS.

- 1) Menentukan simpul akar (*Root Node*)  
Perhitungan nilai entropy dan information gain setiap atribut untuk menentukan simpul akar dihitung menggunakan persamaan 2 dan 3. Atribut yang memiliki nilai *information gain* paling tinggi dipilih sebagai simpul akar.

Tabel 2. Keterangan atribut yang digunakan

Atribut	Nilai Information Gain
Gender	0,172
Minat	0,054
Rata-rata IPA	0,207
Rata-rata IPS	0,023
Psikotes IPA	0,003
Psikotes IPS	0,009
Asal sekolah	0,011

Berdasarkan tabel 2 dapat diketahui bahwa atribut Rata-rata IPA memiliki nilai *gain* yang tertinggi. Sehingga dipilih sebagai simpul akar.

- 2) Menentukan internal node pertama  
Penentuan internal node pada cabang rata-rata IPA  $70 < \text{nilai} \leq 80$  didapatkan nilai *information gain* seperti pada tabel 3. Atribut yang memiliki nilai *information gain* yang tertinggi dipilih sebagai *internal node*.

Tabel 3. Nilai *information gain* penentuan atribut sebagai *internal node* pertama

Atribut	Nilai Information Gain Cabang Rata-rata IPA $70 < \text{nilai} \leq 80$
Gender	0,170
Minat	0,063
Rata-rata IPS	0,220
Psikotes IPA	0,029
Psikotes IPS	0,034
Asal sekolah	0,040

Hasil perhitungan dalam tabel 3 dapat disimpulkan bahwa atribut Rata-rata IPS merupakan *internal node* pada rata-rata IPA  $70 < \text{nilai} \leq 80$  karena memiliki nilai *gain* yang tertinggi dibandingkan dengan atribut yang lain.

- 3) Menentukan internal node kedua  
Menentukan *internal node* kedua pada rata-rata IPA  $70 < \text{nilai} \leq 80$  dan rata-rata IPS  $70 < \text{nilai} \leq 80$  didapatkan nilai *information gain* seperti pada tabel 4.  
Nilai *information gain* dalam tabel 4 menunjukkan bahwa atribut Gender merupakan *internal node* pada cabang rata-rata IPA dan rata-rata IPS  $70 < \text{nilai} \leq 80$  karena memiliki nilai *gain* yang tertinggi dibandingkan dengan atribut yang lain.

Tabel 4. Nilai *information gain* penentuan atribut sebagai *internal node* kedua

Atribut	Nilai Information Gain Cabang Rata-rata IPA $70 < \text{nilai} \leq 80$ dan Rata-rata IPS $70 < \text{nilai} \leq 80$
Gender	0,386
Minat	0,245
Psikotes IPA	0,065
Psikotes IPS	0,060
Asal sekolah	0,114

- 4) Menentukan leaf node  
Menentukan *leaf node* pada rata-rata IPA dan IPS  $70 < \text{nilai} \leq 80$  dengan *gender* laki-laki.

**Tabel 5.** Nilai *information gain* penentuan atribut sebagai *internal node* berikutnya

Atribut	Nilai Information Gain Cabang Rata-rata IPA $70 < \text{nilai} \leq 80$ dan Rata-rata IPS $70 < \text{nilai} \leq 80$ , dan Gender Laki-laki
Minat	0,000
Psikotes IPA	0,000
Psikotes IPS	0,000
Asal sekolah	0,000

Dari hasil dalam tabel 5 dapat disimpulkan bahwa *gender* laki-laki menghasilkan *leaf node* jurusan IPS, dikarenakan hasil dari semua *information gain* bernilai 0, sementara probabilitas siswa jurusan IPA dalam data tidak ada.

#### 4.2 PERHITUNGAN ALGORITMA NAÏVE BAYES

Menggunakan data pelatihan yang ada dan dengan menerapkan metode *naïve bayes* maka dapat dilakukan prediksi terhadap data uji. Sebagai contoh adalah jika terdapat siswa dengan data sebagai berikut: *gender* laki-laki, *minat* IPS, *nilai rata-rata IPA*  $70 < \text{nilai} \leq 80$ , *nilai rata-rata IPS*  $70 < \text{nilai} \leq 80$ , *nilai psikotest IPA*  $25 < \text{skor} \leq 45$ , *nilai psikotest IPS*  $\text{skor} > 65$ , *asal sekolah* wilayah II, maka jurusan siswa tersebut dapat diprediksi. Tabel 6 menunjukkan nilai *confidence* atau nilai HMAP (*Hypothesis Maximum Apriori Probability*) yang dapat digunakan sebagai dasar melakukan prediksi jurusan siswa.

**Tabel 6.** Nilai *confidence*/HMAP terhadap data uji

Probability	Nilai Confidence (HMAP)
P (X1 = Laki-laki, X2 = IPS, X3 = $70 < \text{nilai} \leq 80$ , X4 = $70 < \text{nilai} \leq 80$ , X5 = $25 < \text{Skor} \leq 45$ , X6 = $\text{Skor} > 65$ , X7 = Wilayah II   Y = IPA)	0,000
P (X1 = Laki-laki, X2 = IPS, X3 = $70 < \text{nilai} \leq 80$ , X4 = $70 < \text{nilai} \leq 80$ , X5 = $25 < \text{Skor} \leq 45$ , X6 = $\text{Skor} > 65$ , X7 = Wilayah II   Y = IPS)	0,000

Berdasarkan nilai HMAP yang ditunjukkan pada tabel 6 untuk masing-masing *probability* terhadap jurusan, maka dapat dihasilkan nilai prediksi adalah jurusan IPS. Hal ini dapat diketahui berdasarkan nilai HMAP untuk jurusan IPS yaitu sebesar 0,0000542 yang lebih besar dibandingkan dengan nilai HMAP untuk jurusan IPA yaitu sebesar 0,00000117.

#### 4.3 PERHITUNGAN ALGORITMA K-MEANS

Algoritma K-Means digunakan untuk mengetahui pola pengelompokan jurusan siswa berdasarkan nilai *Distance Performance* terhadap variabel-variabel yang diajukan. *Distance performance* dalam metode klastering dapat dihitung jika nilai pada setiap variabel memiliki tipe numerik. Sehingga data yang digunakan dalam proses ini adalah data numerik.

Pada metode ini, pengelompokan data penjurusan siswa dilakukan perhitungan dengan beberapa tahap yaitu:

- a. Menentukan jumlah kluster  
Kluster yang diinginkan dalam proses ini ditentukan sebanyak 5 kluster (nilai k).
- b. Menentukan nilai *centroid*

Tabel 6 adalah nilai *centroid* dari masing-masing variabel independen pada masing-masing kelompok data.

**Tabel 7.** Nilai *centroid* tiap kluster

Cluster	Nilai Centroid Tiap Variabel						
	X1	X2	X3	X4	X5	X6	X7
1-60	7	9,2	10	11,6	12,4	18	7,2
61-120	7,2	10,2	10,6	11,6	11,8	19	12,2
121-180	6,8	9,2	9,2	11	14,6	21,2	13,2
181-240	7,2	9,4	9,8	12	12,6	17,4	10,4
241-302	8,8	10	7,2	11	13,6	18,6	8,8

Tahap berikutnya adalah mencari jarak antar data untuk melakukan pengelompokan data dengan menggunakan persamaan 5. Hasil perhitungan digunakan untuk mencari nilai jarak *Euclidean*. Suatu data yang memiliki jarak terdekat dengan data lain maka dapat dikelompokkan menjadi satu kluster. Sehingga pada tahap ini menghasilkan sebuah model kluster (*cluster model*) untuk mengetahui kelompok-kelompok jurusan siswa berdasarkan variabel-variabel bebas yang diajukan. Gambar 3 adalah model kluster yang terbentuk.

#### 4.4 PERBANDINGAN TIGA METODE DI ATAS

Perbandingan ketiga metode diperlukan dalam penelitian ini untuk mengetahui metode yang paling baik berdasarkan kriteria nilai *Accuracy*, *Precision* dan *Recall*. Tabel 8 adalah nilai masing-masing kriteria pada 3 metode yang digunakan.

Berdasarkan hasil perbandingan pada tabel 8 dapat disimpulkan bahwa metode *Decision Tree* lebih baik digunakan untuk penelitian ini dikarenakan memiliki nilai *accuracy* dan *recall* yang lebih tinggi dibandingkan dengan metode lainnya. Sedangkan metode *naïve bayes* memiliki nilai *precision* yang lebih tinggi dibandingkan dengan metode yang lain.

**Tabel 8.** Perbandingan 3 metode berdasarkan *accuracy*, *precision* dan *recall*

Komponen	Nilai Accuracy	Nilai Precision	Nilai Recall
Decision Tree	79,14%	75,51%	90,80%
Naïve Bayes	76,82%	77,51%	80,37%
K-Means	36,40%	64,25%	25,40%

### 5. KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan maka dapat disimpulkan telah diperoleh klasifikasi penjurusan siswa menggunakan metode *decision tree*. Hasil klasifikasi menunjukkan bahwa variabel yang paling tinggi

pengaruhnya terhadap penjurusan siswa adalah nilai rata-rata IPA. Hal ini dibuktikan dengan variabel nilai rata-rata IPA menempati sebagai simpul akar pada diagram pohon keputusan.

Proses klustering telah menghasilkan 5 kelompok klaster yang terdiri dari klaster 0 sampai dengan klaster 4 yang masing-masing terdiri dari 24, 53, 16, 109, dan 100 siswa dengan total 302 data siswa. Klaster yang memiliki anggota terbanyak adalah klaster 3, sedangkan yang memiliki anggota paling sedikit adalah klaster 2.

Berdasarkan nilai *accuracy* dan *recall*, metode *decision tree* lebih baik dibandingkan dengan metode yang lain dengan nilai *accuracy* sebesar 79,14% dan nilai *recall* sebesar 90,80%. Berdasarkan nilai *precision*, metode *naive bayes* lebih baik dibandingkan dengan metode yang lain dengan nilai 77,51%.

## DAFTAR PUSTAKA

- [1] Y. S. Nugroho and Setyawan, Klasifikasi Masa Studi Mahasiswa Fakultas Komunikasi dan Informatika, Jurnal Komunikasi dan Teknologi Informasi (KomuniTi) ISSN: 2087-085X, Volume VI No. I Maret, 2014.
- [2] P. Bühlman and B. Yu, Analyzing Bagging, The Annals of Statistics, Vol. 30 no. 4, hal 927-961, 2002.
- [3] D. Kiron, R. Shockley, N. Kruschwitz, G. Finch, and M. Haydock, Analytics: The Widening Divide, MIT Sloan Management Review, 53(2), 1-22, 2012.
- [4] D. Anggraini, Analisis Perubahan Kelompok Berdasarkan Perubahan Nilai Jual Pada Bloomberg Market Data dengan Menggunakan Formal Concept Analysis, Report Paper, 2009.
- [5] Y. S. Nugroho, Analisis Faktor-Faktor Yang Mempengaruhi Tingkat Daya Beli Konsumen Terhadap Listrik Pada Sektor Rumah Tangga: Studi Kasus Kota Salatiga. Thesis, Universitas Gadjah Mada, 2009.
- [6] Ranny dan Budi, Pemilihan Diet Nutrien bagi Penderita Hipertensi Menggunakan Metode Klasifikasi Decision Tree, Jurnal Teknik ITS, Vol. 1, No.1, 2012.
- [7] Bustami, Penerapan Algoritma Naive Bayes untuk Mengklasifikasi Data Nasabah Asuransi. Jurnal Penelitian Teknik Informatika, 2013.
- [8] Y. Agusta, K-Means: Penerapan Permasalahan dan Metode Terkait, Jurnal Sistem dan Informatika Vol.3 hal : 47-60, 2007.