

# ***ANALISIS KOMPARASI ALGORITMA KLASIFIKASI DATA MINING UNTUK PREDIKSI PENJURUSAN SISWA SEKOLAH MENENGAH ATAS (SMA) PRAMITA KARAWACI TANGERANG***

Henny Yulianti

Universitas Pramita Indonesia Fakultas Teknik, Jurusan Teknik Informatika  
Jl. Kampus Pramita Curug Tangerang  
e-mail: hyulia.999@gmail.com

## Abstrak

Penjurusan merupakan salah satu faktor penting di Sekolah Menengah Atas (SMA). Dikarenakan akan mempengaruhi hasil belajar dan menyangkut karir dari peserta didik dimasa depan. Oleh karena itu dibutuhkan metode untuk menangani permasalahan tersebut. Data mining merupakan metode untuk mengolah data dari data yang tidak mempunyai arti menjadi sebuah informasi yang berguna. Salah satu algoritma data mining adalah algoritma klasifikasi. Berbagai algoritma klasifikasi yang ada bisa digunakan untuk mengukur tingkat akurasi. Penelitian ini menggunakan algoritma *decision tree* dan *naïve bayes*. Data yang digunakan sebanyak 365 peserta didik di SMA Pramita Karawaci Tangerang. Hasil dari penelitian untuk mengetahui algoritma klasifikasi yang paling akurat untuk menentukan penjurusan peserta didik SMA.

Kata kunci: Peserta didik, *decision tree*, *naïve bayes*, akurasi

## Abstract

*Majors are one of the important factors in high school (SMA). Because it will affect learning outcomes and concerns the careers of students in the future. Therefore, we need a method to deal with these problems. Data mining is a method for processing data from data that has no meaning into useful information. One data mining algorithm is a classification algorithm. Various classification algorithms that can be used to measure the level of accuracy. This research uses decision tree and naïve bayes algorithms. The data used were 365 students at Pramita Karawaci Tangerang High School. The results of the study to find out the most accurate classification algorithm for determining the direction of high school students.*

*Keywords: students, decision tree, naïve bayes, accuracy*

## I. PENDAHULUAN

Penjurusan peserta didik merupakan salah satu faktor penting dalam kegiatan belajar di Sekolah Menengah Atas (SMA). Salah satu penjurusan akan menyebabkan peserta didik kurang semangat dalam mengikuti kegiatan belajar akibatnya mempengaruhi hasil belajar dan karir di masa depan. Pihak akademik perlu cermat dan perlu penanganan khusus untuk menentukan penjurusan peserta didik yang terarah dan tepat sasaran. Penjurusan di Sekolah Menengah Atas memiliki penjurusan yang berbeda beda tergantung dari pihak sekolah. Biasanya dipengaruhi oleh sarana dan prasarana yang tersedia. Penjurusan tersebut antara lain jurusan Matematika dan Ilmu Pengetahuan Alam (MIPA), Ilmu Bahasa, dan Ilmu Pengetahuan Sosial (IPS)[1]. Untuk MIPA mempunyai mata pelajaran inti antara lain: Matematika, Biologi, Fisika dan Kimia. Mata pelajaran untuk Ilmu Bahasa adalah: Bahasa Indonesia dan Bahasa Inggris. Sedangkan IPS adalah Sejarah, Ekonomi, Geografi dan Sosiologi. Mengingat pentingnya penjurusan untuk meminimalisir dampak yang timbul dari kesalahan penjurusan diperlukan sebuah metode untuk menangani permasalahan tersebut. Salah satu metode untuk menanganinya adalah Data Mining. Data Mining merupakan sebuah metode untuk pengolahan data. Data diolah sehingga menghasilkan informasi [2]. Berbagai algoritma data mining yang ada seperti algoritma klasifikasi sangat cocok untuk diterapkan pada permasalahan penentuan penjurusan peserta didik tingkat Sekolah Menengah Atas. Algoritma Klasifikasi diantaranya Naïve Bayes dan Decision Tree. Alasan peneliti menggunakan naïve bayes

dan decision tree dikarena banyak peneliti menerapkannya diberbagai bidang. Seperti penelitian yang dilakukan oleh Shiju Sathyadevan [3] menerapkan naïve bayes untuk meningkatkan dokumen klasifikasi. Jitendra Agrawal [4] menerapkan naïve bayes untuk memodifikasi kepadatan berdasarkan spasial pengelompokan algoritma yang mempunyai database besar. Sedangkan untuk decision tree diantaranya penelitian yang dilakukan oleh Thangaparvathi [5] menghasilkan sebuah keputusan yang Cepat dalam mengimplementasikan decision tree untuk mengolah dataset besar / database besar. Kwather Aldhah [6], menerapkan decision tree untuk meningkatkan pengetahuan ekstraksi hadis. Adapula penelitian yang melibatkan dua algoritma klasifikasi naïve bayes dan decision tree, seperti penelitian yang dilakukan oleh Sa'diyah Noor Novita Alfisahrin [7], mengoptimalisasi klasifikasi penyakit hati. Dalam penelitian ini, dilakukan analisis dari kedua algoritma untuk mengetahui tingkat akurasi yang baik sehingga bisa digunakan untuk penentuan penjurusan siswa tingkat Sekolah Menengah Atas.

## II. METODE PENELITIAN

### 1. DECISION TREE / C4.5

Decision Tree adalah salah satu teknik yang paling banyak digunakan untuk klasifikasi. Dikarenakan mempunyai akurasi yang kompetitif dan sangat efisien. Ciri khas dari decision tree adalah model direpresentasikan sebagai pohon (pohon keputusan)



**Gambar 1. Contoh Pohon Keputusan**

Aturan yang sering digunakan dalam Decision Tree adalah Entropy. Entropy (D) memiliki persamaan sebagai berikut.

$$Entropy(D) = \sum_{i=1}^n -p_i \log_2 p_i \dots\dots (1)$$

Dimana:

D : Himpunan kasus

A : Atribut

n : jumlah partisi D

pi : proporsi dari Di terhadap D

**2. NAÏVE BAYES**

Naïve Bayes merupakan salah satu algoritma klasifikasi yang banyak digunakan untuk mengklasifikasi text. Secara fundamental naïve bayes menggunakan metode probabilitas dan statistika. Metode ini digunakan untuk memprediksi kemungkinan / probabilitas di masa yang akan datang dengan berpedoman pada masa sebelumnya. Naïve Bayes menghasilkan model yang sangat akurat. Model seketika dapat diperbarui dengan mudah ketika adanya tambahan data baru. Hal itu tidak terlepas dari peranan probabilitas. Adapun rumus naïve bayes dapat dirumuskan sebagai berikut :

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)} \dots\dots\dots (2)$$

Keterangan :

H: Hipotesis data X merupakan suatu class spesifik

X : Data dengan class yang belum diketahui

P(H) : Probabilitas hipotesis H (prior probability)

P(H|X): Probabilitas hipotesis H berdasar kondisi X (posteriori probability)

P(X|H): Probabilitas X berdasarkan kondisi pada hipotesis H

P(X) : Probabilitas X

**3. DATASET**

Data penelitian diperoleh dari SMA mempunyai 365 data peserta didik kelas X pada tahun ajaran 2014/2015 dengan berbagai atribut antara lain sebagai berikut.

Table – 1. Tabel Atribut

No.	Nama Atribut
1.	Nomor Pendaftaran
2.	NISN
3.	Nama
4.	Jenis Kelamin
5.	Tempat Lahir
6.	Tanggal Lahir
7.	Agama
8.	Asal Sekolah
9.	Nomor Pendaftaran
10.	Alamat
11.	Nama Orang Tua
12.	Nomor Telepon Rumah
13.	Nomor Handphone
14.	Pekerjaan Orang Tua
15.	Nilai Bahasa Indonesia
16.	Nilai Bahasa Inggris
17.	Nilai Matematika
18.	Nilai IPA

19.	Total Nilai
20.	Minat
21.	PL
22.	Keterangan
23.	Seragam
24.	Jurusan

Data diolah sudah mengalami pre-processing dan tidak adanya missing value sehingga bisa dilakukan komprasi data.

Pada kurikulum 2013, skala penilaian tidak lagi pada rentang nilai 0-100 melainkan dikonversikan antara 1-4 untuk aspek psikomotor dan kognitif, sedangkan aspek afektif menggunakan SB='Sangat baik', B='Baik', C='Cukup', K='Kurang'.

Tabel – 2. Konversi Nilai

Nilai	Predikat
3,33 – 4	1
2,33 – 3,32	2
1,33 – 2,32	3
0 – 1,32	4

Metode yang digunakan pada penelitian ini menggunakan berbagai proses. Mulai dari pengumpulan data, pre-processing proses klasifikasi dan evaluasi dari algoritma yang digunakan untuk penelitian.

#### 1. Pengumpulan data.

Pada tahapan ini proses pengumpulan data. Dan sebagai data objek peserta didik SMA Pramita Karawaci Tangerang. Peneliti melakukan pemilihan data terlebih dahulu sebelum ke tahapan berikutnya. Data diperoleh nantinya digunakan untuk penentuan penjurusan dengan menggunakan algoritma klasifikasi.

#### 2. Pre-Processing( Integrasi, Selection, Cleansing),

Setelah data dirasa cocok, tahapan berikutnya melakukan pemrosesan data

apakah ada data yang kosong atau data yang cocok. Hasil dari tahapan ini data bisa digunakan ke tahap klasifikasi.

#### 3. Proses Klasifikasi

Proses mengolah data dengan berbagai algoritma antara lain decision tree dan naïve bayes. Data diolah dengan data training dan testing yang berbeda agar bisa di evaluasi dan dianalisis.

#### 4. Evaluasi

Tahapan ini melakukan proses analisa atau evaluasi dari algoritma decision tree dan naïve bayes dari segi accuracy, precision, recall, mean square error.

### III. HASIL DAN PEMBAHASAN

Hasil penelitian dibedakan menjadi 3 jenis yaitu hasil pengujian Decision Tree, Naïve Bayes, dan Analisis dari kedua algoritma tersebut. Penelitian ini menggunakan teknik split validation dengan membagi data dalam dua kelompok yaitu Data Training dan Data Testing. Data Training dan Data Testing dengan besaran atau perbandingan 90%:10 %, 80%:20%, 70%:30, 60%:40%.

A: Testing 90% : Training 10%

B: Testing 80% : Training 20%

C: Testing 70% : Training 30%

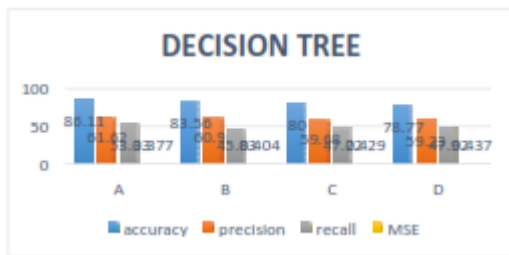
D: Testing 60% : Training 40%

#### 1. DECISION TREE

Berdasarkan hasil penelitian dengan 4 kali prosentasi data training dan testing yang berbeda didapatkan hasil sebagai berikut.

Tabel – 3. Tabel Hasil Decision Tree

	A	B	C	D	Avg
<i>Accuracy</i>	86,1	83,5	80,0	78,7	82,1
	1%	6%	0%	7%	1%
<i>Precision</i>	61,6	60,9	59,6	59,2	60,3
	2%	5%	8%	3%	7%
<i>Recall</i>	53,3	45,8	47,2	47,9	48,5
	3%	3%	2%	2%	8%
<i>MSE</i>	0.37	0.40	0.42	0.43	0.41
	7	4	9	7	2



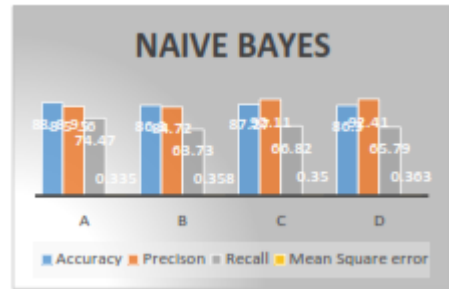
Gambar 2 Diagram Accuracy, Precision, Recall Decision Tree

## 2. NAÏVE BAYES

Berdasarkan hasil penelitian dengan 4 kali prosentasi data training dan testing yang berbeda didapatkan hasil sebagai berikut.

Tabel – 4. Tabel Hasil Algoritma Naïve Bayes

	A	B	C	D	Avg
<i>Accuracy</i>	88,8	86,3	87,2	86,3	87,1
	9%	0%	7%	0%	9%
<i>Precision</i>	85,5	84,7	92,1	92,4	88,7
	6%	2%	1%	1%	%
<i>Recall</i>	74,4	63,7	66,8	65,7	67,7
	7%	3%	2%	9%	0%
<i>MSE</i>	0.33	0.35	0.35	0.36	0.35
	5	8	0	3	2



Gambar 3 Diagram Accuracy, Precision, Recall, dan MSE Naïve Bayes

## 3 ANALISIS PENGUJIAN

Berdasarkan hasil penelitian Decision Tree dan Naïve Bayes diperoleh perbandingan sebagai berikut.

Tabel – 5. Perbandingan Decision Tree dan Naïve Bayes

	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>MS E</i>
<i>Decision Tree</i>	82,11%	60,37%	48.58 %	0.41 2
<i>Naïve Bayes</i>	87,19%	88,7%	67.70 %	0.35 2

Dari tabel diatas Naïve Bayes memiliki tingkat akurasi yang baik dibanding dengan decision tree dengan 87,19% : 82,11 %. Begitu pula untuk Precision , Recall dan MSE. Dengan kata lain Naïve Bayes lebih baik dibandingkan dengan decision tree.

## IV. KESIMPULAN DAN SARAN

Dari ujicoba dengan berbagai komposisi data training dan testing dapat disimpulkan bahwa algoritma *Naïve Bayes* menghasilkan tingkat akurasi sangat baik yaitu mencapai 82,11 % dibandingkan dengan algoritma decision tree. Sehingga Naïve bayes sangat cocok diterapkan untuk periksi penjurusan siswa pada jenjang sekolah menengah keatas.

Adapun saran untuk penelitian kedepannya dengan membandingkan algoritma klasifikasi lainnya dan menambahkan atribut seperti penghasilan orang tua, lingkungan sekitar dan sebagainya.

### **DAFTAR PUSTAKA**

- [1] Departemen Pendidikan Nasional, Panduan Penyusunan Laporan Hasil Belajar Peserta Didik Sekolah Menengah Atas (SMA). Jakarta, Indonesia, 2006.
- [2] Witten Ian H., Eibe Frank, and Hall Mark A., Data Mining : Practical Machine Learning Tools and Techniques 3rd Edition. Elsevier, 2011.
- [3] Shiju Sathyadevan, "Improved Document Classification Through", International Conference on Data Science & Engineering (JCDSE), 2014
- [4] Jitendra Agrawal, "Modification of Density Based Spatial Clustering Algoritn for Large Database Using Naive's Baye's Theorem", Fourth International Conference on Communication Systems and Network Technologies, 2014
- [5] Thangaparvathi, "A High Speed Decision Tree Classifier Algorithm for Huge Dataset", IEEE-International Conference on Recent Trends in Information Technology, 2011