# Analisis Komparasi Algoritma Naive Bayes Dan C4-5 Untuk Waktu Kelulusan Mahasiswa

**Risqiati[(1)], Bambang Ismanto[(2)]**
*STMIK Widya Pratama Pekalongan*
Jl. Patriot 25 Pekalongan Telp (0285) 427816
[(1)] **e***mail:risqiati24@gmail.com*
[(2)] *email:* bams.stmikwp@gmail.com

***ABSTRAK***

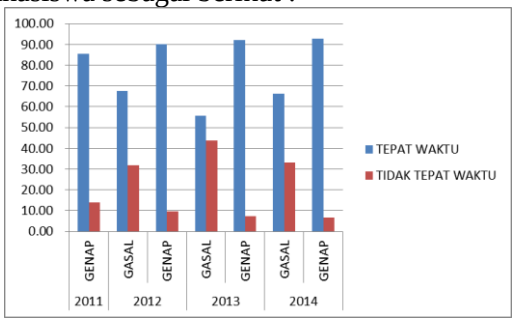
*STMIK Widya Pratama merupakan salah satu perguruan tinggi swasta yang ada di Pekalongan. Mahasiswa merupakan indicator maju atau tidak nya suatu institusi pendidikan. Sedikit banyaknya mahasiswa yang tidak tepat waktu kelulusannya berdampak pada akreditasi sebuah Perguruan Tinggi Swasta. Penelitian ini bertujuan untuk mengkomparasi data kelulusan mahasiswa STMIK Widya Pratama dari tahun 2011 – 2014 menggunakan dua algoritma yaitu algoritma Naïve Bayes dengan algoritma C4.5, sehingga dapat mengklasifikasi mahasiswa lulus tepat waktu dengan akurasi yang baik. Proses klasifikasinya menggunakan algoritma naïve bayes yang merupakan teknik prediksi berbasis probabilistik sederhana yang modelnya mengunakan fitur independen. Independen yang dimaksud adalah bahwa pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Sedangkan algoritma C4.5 yang menghasilkan pohon keputusan akan bekerja dengan baik bila bias yang ada sedikit. Implementasi menggunakan Rapid Miner 5.3 digunakan untuk membantu menemukan nilai yang akurat.*

***Kata Kunci*** : *Data mining, Komparasi, Klasifikasi,Naïve Bayes, C4.5*

## 1. Pendahuluan

### 1.1 Latar Belakang

Pendidikan merupakan hal yang sangat penting dalam kehidupan manusia. Dewasa ini, lembaga pendidikan tumbuh dan berkembang dalam rangka pemenuhan kebutuhan primer. Dalam institusi pendidikan, ada beberapa faktor yang digunakan sebagai indikator maju atau tidaknya institusi pendidikan tersebut, antara lain: dosen, sarana prasarana dan mahasiswa. Mahasiswa adalah salah satu aspek penting dalam evaluasi keberhasilan penyelenggaraan program studi pada suatu perguruan tinggi, karena sekarang mahasiswa bukan lagi objek, melainkan subjek yang memerankan peranan penting dalam kehidupan kampus. Pemantauan mahasiswa yang masuk ke perguruan tinggi,, peningkatan kemampuan mahasiswa yang diperoleh, prestasi yang dicapai serta rasio kelulusan terhadap jumlah total mahasiswa seyogyanya mendapatkan perhatian yang serius untuk memperoleh kepercayaan perusahaan dalam menilai dan menetapkan penggunaan kelulusannya (Mujib Ridwan, Hadi Suyono, dan M. Sarosa 2013). Kelulusan mahasiswa yang tepat waktu, menjadi salah satu indikator keberhasilan institusi pendidikan dalam mengelola sumber daya mahasiswa. Data dari unit sumber daya informasi (USDI) STMIK Widya Pratama didapatkan hasil kelulusan mahasiswa sebagai berikut :




Gambar 1 Data Jumlah Kelulusan Mahasiswa Selama 4 Tahun

Dari gambar 1 diketahui adanya perbedaan jumlah yang cukup banyak antara kelulusan mahasiswa pada semester gasal dan genap. Penurunan jumlah mahasiswa yang lulus tidak tepat waktu dalam satu tahun sangat berbeda. Bagi STMIK Widya Pratama kondisi tersebut tidaklah kondusif, karena salah satu syarat untuk akreditasi sekolah tinggi adalah rata - rata masa studi dan rata - rata IPK lulusan (BAN PT, 2008).

Beberapa algoritma klasifikasi data mining telah digunakan untuk memprediksi kelulusan mahasiswa diantaranya algoritma naïve bayes, chaid, algoritma C4-5. Hasilnya, uji coba naïve bayes diperoleh tingkat kesalahan prediksi

sebesar 20% - 50%. Sedangkan menggunakan chaid didapatkan hasil 88,6% untuk kelulusan tepat waktu dan 99% untuk kelulusan tidak tepat waktu berdasarkan jurusan yang berbeda. Untuk C4-5 didapatkan hasil bahwa beberapa variabel menjadi penentu probabilitas tertinggi untuk kelulusan.

Dalam penelitian ini, dilakukan analisis komparasi 2 algoritma yaitu naïve bayes dan C4-5 menggunakan data kelulusan mahasiswa dari tahun 2011 – 2014 dengan jumlah 12 variabel dan 1076 record.

## 1.2 Landasan Teori

### 1.2.1 Klasifikasi

Klasifikasi merupakan suatu kegiatan untuk menilai objek data untuk memasukannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua jenis pekerjaan utama yang harus dilakukan, yaitu (1) pembangunan model sebagai prototype akan disimpan sebagai memori dan (2) menggunakan model yang ada untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya (Eko Prasetyo, 2012). Adapun algoritma yang termasuk ke dalam metode klasifikasi antara lain C4.5, *Rain Forest, Naïve Bayesian, neural network, genetic algorithm, fuzzy, case-based reasoning, dan k-Nearest Neighbor* (Nobertus Krisandi, et., al., 2013).

### 1.2.2 Naïve Bayes

Naive bayes adalah bentuk sederhana dari pengkalsifikasi jaringan Bayesian. Hal ini karena asumsi kemandirian kondisi dalam naive bayes relatif benar dalam kenyaataannya, terutama dalam hal atribut yang kompleks. Dalam hal klasifikasi akurasi atau tingkat kesalahan, naive bayes tetap menjaga efisiensi dan kesederhanaan (Levent Koc, 2012) (Liangxiao Jiang, et., al., 2012). Naive bayes juga merupakan metode yang sederhana dalam hal yang berbasis probabilitas yang dapat memprediksi keanggotaan kelas. Naive Bayes memiliki keuntungan (a) mudah digunakan, dan (b) hanya satu scan data pelatihan yang diperlukan untuk generasi probabilitas (Dewan Md. Farid, 2014). Bentuk umum dari Naive Bayes sebagai berikut (Eko Prasetyo, 2012):

$$P(H \mid E) = \frac{P(E \mid H) x P(H)}{P(E)}$$

Keterangan :

P (H | E) = Probabilitas akhir bersyarat (*conditional probability*) suatu hipotesis H terjadi jika diberikan bukti (*evidence*) E terjadi

P (E | H) = Probalitas sebuah bukti E terjadi akan mempengaruhi hipotesis H

P (H) = Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apa pun

P (E) = Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis/bukti yang lain.

### 1.2.3 C4-5

C4.5 Merupakan pengembangan dari algoritma ID3 (Larose 2005) yang dikembangkan oleh Quinlan (Han and Kamber 2006). Algoritma C4.5 banyak digunakan peneliti untuk melakukan tugas klasifikasi. *Output* dari algoritma C4.5 adalah sebuah pohon keputusan atau sering dikenal dengan *decissin tree.*

Langkah untuk membuat sebuah *decision tree* dari algoritma C4.5 adalah sebagai berikut (Han and Kamber 2006b):

1. 1. Mempersiapkan data training, *data training* yaitu data yang diambil dari data histori yang pernah terjadi sebelumnya atau disebut data masa lalu dan sudah dikelompokan dalam kelas-kelas tertentu.
   2. Menentukan akar pohon. Akar pohon ditentukan dengan cara menghitung *GainRatio* tertinggi dari masing-masing atribut. Sebelum menghitung *GainRatio*, terlebih dahulu menghitung *Total Entropy* sebelum dicari masing-masing *Entropy class*, adapun rumus mencari Entropy seperti di bawah:

$$Entropy(S) = \sum_{i=1}^{n} - p * \log 2 pi$$

Keterangan:
$S$ = Himpunan kasus
$n$ = jumlah partisi $S$
$pi$ = proporsi dari $S_i$ terhadap $S$

Dimana $log_2 pi$ dapat dihitung dengan cara:
1. Menghitung nilai *GainRatio* sebagai akar pohon, tetapi sebelumnya menghitung *Gain* dan *SplitEntropy (SplitInfo)*, rumus untuk menghitung *Gain* seperti dibawah ini: Rumus untuk menghitung *GainRatio*, dibawah ini:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$

Keterangan:
$S$ = Himpunan Kasus
$A$ = Atribut
$n$ = jumlah partisi atribut $A$
$|S_i|$ = jumlah kasus pada partisi ke-$i$
$|S|$ = jumlah kasus dalam $S$
3. Ulangi langkah ke-2 dan ke-3 hingga semua tupel terpartisi
4. Proses partisi pohon keputusan akan berhenti disaat:
   a. Semua tupel dalam node $N$ mendapatkan kelas yang sama
   b. Tidak ada atribut didalam tupel yang dipartisi lagi
   c. Tidak ada tupel didalam cabang yang kosong

## 2. Metode Penelitian

Tahap pertama yang dilakukan adalah pengumpulan data. Data yang diperoleh sejumlah 1076 *record* dari wisuda 4 tahun terakhir. Dengan rincian data tahun 2011 pada bulan September sejumlah 99 *record*, data tahun 2012 bulan April sejumlah 126 *record*, data tahun 2012 bulan Oktober sejumlah 130 *record*, data tahun 2013 bu lan April sejumlah 125 *record*, data tahun 2013 bulan September sejumlah 205 *record*, data tahun 2014 bulan Maret sejumlah 152 *record*, *dan* data tahun 2014 bulan Agustus sejumlah 239 *record*. Dalam data Kelulusan Mahasiswa STMIK Widya Pratama Pekalongan terdapat 12 atribut dan 1 atribut sebagai label yaitu "Tepat Waktu" dan "Tidak Tepat Waktu". Meta data dari data Kelulusan Mahasiswa yang telah digabungkan dari tahun 2011 sampai dengan 2014 dapat dilihat pada tabel 1

Tabel 1 Meta Data Kelulusan Mahasiswa STMIK Widya Pratama

| Atribut | Type | Keterangan |
|---|---|---|
| Jenjang | Binominal | S1, D3 |
| Konsentrasi | Binominal | E-Business, Business Intelligent Sistem, Mobile Application and Web Programming, Multimedia, Computer Grafic and Multimedia, Programming and Database administratio, Komputerisasi Akuntansi |
| Pekerjaan | Binominal | Bekerja, belum bekerja |
| Jns Kel | Binominal | L, P |
| Shift | Binominal | Pagi, malam |
| Biaya | Polinominal | Sendiri, orang tua, beasiswa |
| St sipil | Binominal | Tidak menikah, menikah |
| Ipk | Polinominal | Cukup memuaskan, memuaskan, sangat memuaskan, coumloude |
| Jumlah nilai D | binominal | Nol, satu |
| Jumlah poin | Polinominal | Cukup, bagus, sangat bagus |
| Judul | Polinominal | System informasi, system pendukung keputusan, system administrasi, system pakar, aplikasi bantu, game, animasi, media pembelajaran |
| Keterangan | binominal | Tepat waktu, tidak tepat waktu |

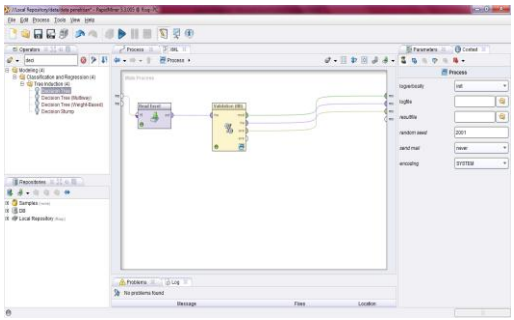
Tahap selanjutnya dilakukan klasifikasi menggunakan algoritma naïve bayes dengan cara melakukan validasi dataset kelulusan mahasiswa untuk diproses secara *training* dan *testing* untuk mendapatkan hasil akurasi.

Setelah itu melakukan klasifikasi algoritma C4-5 dengan cara memilah dataset yang akan menjadi data *training* serta dataset yang nantinya akan menjadi data *testing*. Selanjutnya memilih atribut sebagai atribut akar. Atribut akar dipilih dengan melihat nilai *gain ratio* tertinggi dari semua atribut yang ada. Berikutnya membuat cabang untuk setiap nilai dilanjutkan dengan pembagian kasus dalam tiap cabang. Langkah berikutnya adalah mengulang pembagian cabang untuk tiap nilai sampai dengan semua kasus dalam tiap cabang memiliki kelas yang sama.

Tahap terakhir dilakukan komparasi antara algoritma naïve bayes dan algoritma C4-5 untuk mendapatkan hasil akurasi yang terbaik.

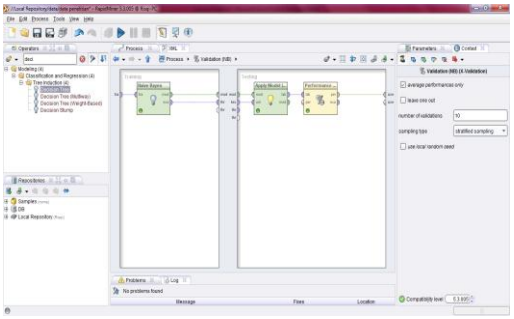
## 3. Hasil dan Pembahasan

Hasil coba klasifikasi algoritma naïve bayes kelulusan mahasiswa. Dalam hal ini menggunakan *X Validation* untuk membantu menghasilkan tingkat keakurasian berdasarkan dataset kelulusan mahasiswa.



Gambar 2 Pemodelan Naive Bayes

Selanjutnya melakukan *training* dan *testing* data. Di kolom *training* terdapat algoritma yang diterapkan yaitu Naïve Bayes. Sedangkan di dalam kolom *testing* terdapat *Apply Model* untuk menjalankan model naïve bayes dan *Performance* untuk mengukur performa dari model Naïve Bayes.



Gambar 3 Proses *Training* dan *Testing* Naive Bayes

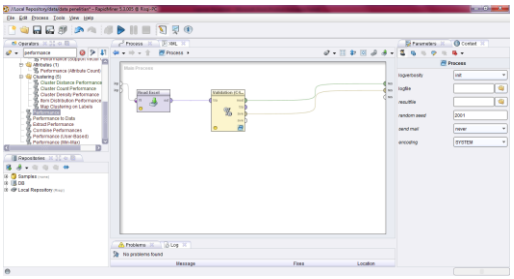
Setelah dilakukan proses *training* dan *testing* data didapatkan hasil akurasi model naïve bayes dengan tingkat akurasi 83.36% yang dievaluasi menggunakan *confusion matrix.*

accuracy: 83.36% +/- 3.14% (mikro: 83.36%)

| | true TEPAT WAKTU | true TIDAKTEPAT WAKTU | class precision |
|---|---|---|---|
| pred. TEPAT WAKTU | 826 | 95 | 89.69% |
| pred. TIDAKTEPAT WAKTU | 84 | 71 | 45.81% |
| class recall | 90.77% | 42.77% | |

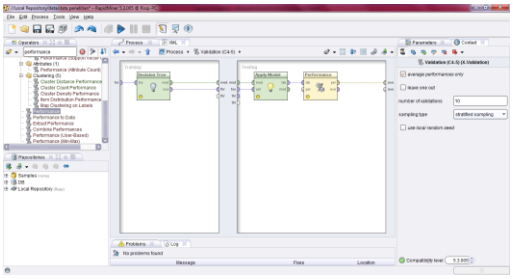
Gambar 4 Hasil Akurasi Algoritma Naïve Bayes

Selanjutnya dilakukan klasifikasi C4-5 dengan cara menggunakan *X Validation* untuk membantu menghasilkan tingkat keakurasian berdasarkan dataset kelulusan mahasiswa.

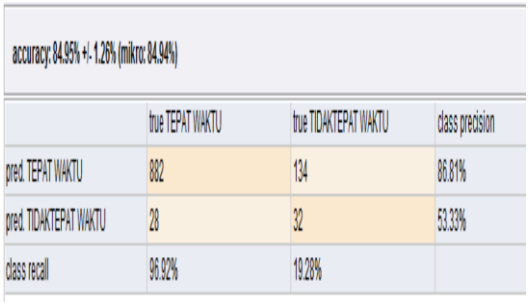


Gambar 5 Pemodelan C4-5

Setelah proses pemodelan selesai dilakukan proses *training* dan *testing* data. Dimana pada kolom *training* terdapat algoritma Decision Tree sebagai algoritma yang digunakan untuk klasifikasi sedangkan pada kolom *testing* terdapat *Apply Model* untuk menjalankan model Decision Tree dan *Performance* untuk mengukur performa dari model Decision Tree.



Setelah proses training dan testing dilakukan maka akan didapatkan hasil akurasi dari Decision Tree dalam hal ini tingkat akurasinya sebesar 84.95%



accuracy: 84.95% +/- 1.26% (mikro: 84.94%)

| | true TEPAT WAKTU | true TIDAKTEPAT WAKTU | class precision |
|---|---|---|---|
| pred. TEPAT WAKTU | 882 | 134 | 86.81% |
| pred. TIDAKTEPAT WAKTU | 28 | 32 | 53.33% |
| class recall | 96.92% | 19.28% | |

Gambar 6 Hasil Akurasi Algoritma C4-5

Proses selanjutnya dilakukan komparasi antara algoritma Naïve Bayes dan algoritma C4-5 untuk mendapatkan hasil akurasi yang terbaik. Adapun untuk mengetahui hasil komparasi hanya memandingkan antara hasil akurasi dari kedua algoritma yang dipakai

Tabel 2 Hasil Komparasi Algoritma

| No. | Algoritma Naïve Bayes (%) | Algoritma C4-5 (%) |
|---|---|---|
| 1 | 83.36 | 84.95 |

Sehingga dapat disimpulkan bahwa algoritma yang baik untuk kelulusan mahasiswa STMIK Widya Pratama adalah algoritma C4-5 dengan selisih hasil akurasi sebesar 1.59%

## 4. Kesimpulan

Dari hasil penelitian dengan dataset mahasiswa STMIK Widya Pratama yang memiliki 12 variabel yang diolah dengan mengunakan software rapid miner versi 5 yang dievaluasi dengan *confusion matrix* , dihasilkan bahwa algoritma C4-5 memiliki nilai akurasi lebih baik dari pada algrotima Naïve Bayes. Nilai yang didapatkan didapatkan hasil komparasi 2 algoritma ini, bahwa algoritma yang bagus hasil akurasinya adalah algoritma C4-5 sebesar 84.95% selisih 1.59% dari algoritma Naïve Bayes.

## 5. Saran dan Penelitian Selanjutnya

Dalam penelitian ini untuk hasil komparasi masih menggunakan dua algoritma yaitu Naïve Bayes dan algoritma C4-5, sehingga memungkinkan untuk menambah algoritma lain supaya akurasi lebih relavan. Dataset yang digunakan masih tergolong dalam dataset mentah dan dalam penelitian berikutnya dataset dapat diolah terlebih dahulu untuk mendapatkan hasil yang lebih baik.

Alangkah baiknya dalam penelitian yang lebih lanjut, diberikan komparasi minimal 4 algoritma klasifikasi dengan data set yang telah diolah. Diharapkan mampu menghasilkan nilai akurasi yang lebih baik dari komparasi 2 algoritma ini.

## Refrensi


Arief Jananto, 2013. Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa. *Jurnal Teknologi Informasi DINAMIK*, Vol. 18 No. 1.

B. Azhagusundari and A. S. Thanamani, 2013. Feature Selection based on Information Gain. pp. pp. 18–21.

BAN PT, 2008. Portofolio Fakultas/Sekolah Tinggi, Akreditasi PS Sarjana.

Budi Santosa, 2007. *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.

Daniel T. Larose, 2005. *Discovery Knowledge in Data : an Introduction to Data Mining*. John Wiley & Sons.

Dewan Md. Farid, et., al., 2014. Hybrid Decision Tree and Naïve Bayes Classifiers for Multi-Class Classification Tasks. *Expert Systems with Applications, 41(4, Part 2)*, pp.pp.1937 - 1946.

E. Alpaydin, 2010. Introduction to Machine Learning Second Edition.

Eko Prasetyo, 2012. *Data Mining Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi Offset.

F. Gorunescu, 2011. *Data Mining: Concept, Models and Techniques*. Berlin: Heidelberg: Springer Berlin Heidelberg.

Girish Chandrashekar and Ferat Sahin, 2014. A Survey On Feature Selection Methods.

I. H. Witten, et., al., 2011. *Data Mining : Practical Machine Learning Tools and Techniques*. 3rd ed. Elsevier.

Ida Ayu Sri Padmini, e..a.., 2012. Analisis Waktu Kelulusan Mahasiswa dengan Metode CHAID (Studi Kasus : FMIPA Universitas UDAYANA). *e-Jurnal Matematika*, Vol. 1, No. 1, pp.pp 89-93.

Kusrini and L. E. Taufiq, 2009. *Algoritma Data Mining*. Yogyakarta: Andi Offset.

Levent Koc, et., al., 2012. A Network Intrusion Detection System Based on a Hidden Naïve Bayes Multiclass Classifier.

Liangxiao Jiang, et., al., 2012. Improving Tree Augmented Naive Bayes for Class Probability Estimation.

Md. Faisal Kabir, et., al., 2011. Enhanced Classification Accuracy on Naive Bayes Data Mining Models. *Journal of Computer Application (0975 - 8887)*, Vol. 28 - No. 3.

Nobertus Krisandi, et., al., 2013. Algoritma k-Nearest Neighbor Dalam Klasifikasi Data Hasil Produksi Kelapa Sawit Pada Pt. Minamas Kecamatan Parindu. *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster*, Vol. 02, No. 1, pp.pp.33 - 38.

Oded Maimon and Lior Rokach, 2010. *Data Mining and Knowledge Discovery Handbook*. 2nd ed. Springer New York Dordrecht Heidelberg London.

Peng-Mian, et., al., 2013. Naïve Bayes Classifier with Feature Selection to Identify Phage Virion Proteins.

Wang, Li-Min, et., al., 2006. Combining decision tree and Naive Bayes for classification. Knowledge-Based Systems.

X. Wu, V. Kumar, et., al., 2007. Top Algorithms in Data Mining. Vol. 14, pp.pp.1-37.

Yusuf Sulistyo Nugroho, 2014. Penerapan Algoritma C4.5 untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi dan Informatika Universitas [23]Muhammadiyah Surakarta. In *In Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 2014*. Yogyakarta, 2014.