Perbandingan Kinerja Algoritma Klasifikasi *Naive Bayes*, *K-Nearest Neighbor*, dan C4.5 Untuk Prediksi Kelulusan Mahasiswa

1st Mohammad Imron , 2nd Satia Angga Kusumah
Program Studi Informatika
Universitas Amikom Purwokerto
Purwokerto, Indonesia
1st imron@amikompurwokerto.ac.id, 2nd chacha.mey17@gmail.com

Abstrak— Perguruan tinggi merupakan salah satu jenjang pendidikan yang dianggap sebagai gerbang terakhir bagi para pelajar untuk menimba ilmu sebelum akhirnya melibatkan diri dalam dunia kerja. Saat ini institusi perguruan tinggi berada dalam lingkungan yang sangat kompetitif. Sehingga perguruan tinggi kini dituntut untuk memiliki keunggulan dalam bersaing dengan memanfaatkan semua sumber daya yang dimiliki. Penelitian yang dilakukan adalah untuk menganalisis mengenai perbandingan metode klasifikasi *Naive Bayes*, *K-Nearest Neighbor*, dan C4.5 untuk menentukan nilai akurasi yang lebih baik, bagaimana menerapkan model prediksi kelulusan mahasiswa STMIK AMIKOM Purwokerto dengan menggunakan pendekatan *data mining*. Hasil pengujian yang sudah dilakukan terhadap *data training* dengan total data 667 dan *data testing 867 record*. Perhitungan *confusion matrix* menunjukkan bahwa algoritma *K-Nearest Neighbor* memiliki tingkat akurasi yang lebih baik dengan prediksi akurasi sebesar 89,04%, dibandingkan dengan algoritma *naive bayes* dan algoritma C4.5

Kata Kunci — Prediksi, *Data Mining*, *Naive Bayes*, *K-Nearest Neighbor*, C4.5

I. LATAR BELAKANG

Perguruan Tinggi merupakan salah satu jenejang pendidikan yang dianggap sebagai gerbang terakhir bagi para pelajar untuk menimba ilmu sebelum akhirnya melibatkan diri dalam dunia kerja. Saat ini institusi Perguruan Tinggi berada dalam lingkungan yang sangat kompetitif, sehingga Perguruan Tinggi kini dituntut untuk memiliki keunggulan dalam bersaing dengan memanfaatkan semua sumber daya yang dimiliki. Selain sumber daya manusia, sarana serta prasaran, sistem informasi adalah contoh lain dari beberapa sumber daya yang dapat digunakan guna meningkatkan kemampuan dan daya saing Perguruan Tinggi. Sistem informasi dalam hal ini dapat digunakan guna memperoleh, mengelola serta menyebarkan informasi yang telah diolah menjadi informasi yan penting bagi Perguruan Tinggi khususnya. Agar dapat menunjang berbagai kegiatan operasional, sekaligus dapat menjadi peran serta dalam pengambilan sebuah keputusan strategis yang akan dilakukan dimasa depan.

Institusi Perguruan Tinggi pastinya harus meningkatkan kualitas layanan dan memuaskan para mahasiswa serta ruang publik disekitar mereka agar mampu bersaing dengan Perguruan Tinggi lainnya. STMIK AMIKOM Purwokerto merupakan salah satu Perguruan Tinggi yang berdiri pada tanggal 16 Mei 2005, dalam peraturan pendidikan tahun 2009 pada BAB I pasa 1 ayat 2 disebutkan bahwa Program Sarjana (S1) reguler adalah program pendidikan akademik setelah pendidikan menengah, yang memiliki beban studi sekurang-kurangnya 144 sks dan sebanyak-banyaknya 160 sks yang dijadwalkan untuk 8 semester [1].

Penelitian tentang kelulusan mahasiswa yang telah diteliti oleh Nugroho tentang penerapan algoritma c4.5 untuk klasifikasi predikat kelulusan mahasiswa fakultas komunikasi informastika Universitas Muhammadiyah Surakarta dengan menggunakan algoritma klasifikasi C4.5 untuk memprediksi kelulusan mahasiswa dengan hasil akurasi prediksi 73,91% [2].

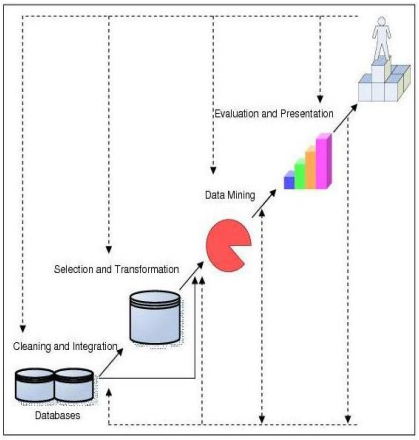
Penelitian yang sudah dilakukan Fiastyanto dalam hal ini memprediksi tingkat kelulusan mahasiswa dengan melakukan perbandingan metode *data mining* yaitu algoritma C4.5 dengan *Naive Bayes* yang diterapkan pada data mahasiswa strata 1 Fakultas Ilmu Komputer Universitas Dian Nuswantoro tahun 2008 s/d 2009. Berdasarkan hasil pengujian dengan mengukur kinerja kedua algoritma tersebut menggunakan pengujian *confusion matrix*, kemudian diketahui bahwa algoritma C4.5 memiliki nilai akurasi yang lebih baik yaitu sebesar 77,35%, sedangkan *naive bayes* memiliki nilai akurasi 74,09% [3].

Akreditasi perguruan tinggi oleh BAN-PT (Badan Akreditasi Nasional Perguruan Tinggi) merupakan salah satu parameter dalam menentukan mutu perguruan tinggi serta program studi di Indonesia. Pada kenyataannya akreditasi suatu perguruan tinggi dapat menjadi salah satu nilai jual bagi perguruan tersebut yang akan menarik minat dari para calon mahasiswa baru. Banyak pula perusahaan atau instansi yang mamandang bahwa semakin tinggi tingkat akreditasi dari suatu perguruan tinggi mencerminkan kualitas dari para mahasiswanya.

Data mining merupakan proses ektraksi pengetahuan dari data yang besar. Sesuai fungsinya, *data mining* adalah proses pengambilan pengetahuan dari volume data yang besar yang disimpan dalam basis data, data *warehouse*, atau infomasi yang disimpan dalam *repository*[4].

Data mining disebut juga dengan istilah *Knowledge-Discovery in Database* (KDD) yang merupakan proses secara otomatis atas pencarian data didalam sebuah penyimpanan yang memiliki kapasitas besar, untuk mengetahui data tersebut dapat diketahui dengan pola seperti klasifikasi hubungan (association) atau dengan pola pengelompokan (*clustering*).

Dan secara sederhana *data mining* dapat diartikan sebagai proses pengekstrak atau "menggali" pengetahuan yang ada pada sekumpulan data, langkah proses *data mining* seperti terlihat pada gambar 2.



Gambar 1. Langkah Proses *Data Mining*.

Dari sudut pandang yang lain, *data mining* dianggap sebagai satu langkah yang penting didalam proses KDD, proses KDD tersebut terdiri dari langkah-langkah sebagai berikut[5]:

1. *Data Cleaning*, proses menghapus data yang tidak konsisten dan kotor.
2. *Data Integration*, penggabungan beberapa sumber data
3. *Data Selection*, pengambilan data yang akan dipakai dari sumber data
4. *Data Transformation*, proses dimana data ditransformasikan menjadi bentuk yang sesuai untuk diproses dalam *data mining*
5. *Data Mining*, suatu proses yang penting dengan melibatkan metode untuk menghasilkan suatu pola data
6. *Pattern Evaluation*, proses untuk menguji kebenaran dari pola data yang mewakili *knowledge knowledge* yang ada didalam data itu sendiri
7. *Knowledge Presentation*, proses visualisasi dan teknik menyajikan *knowledge* digunakan untuk menampilkan *knowledge* hasil *mining* kepada *user*

A. Algoritma *Naive Bayes*

Bayesian *classification* adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*. Bayesian *clasification* didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan *decesion tree* dan *neural network*. Bayesian *classification* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* dengan data yang besar[6].

Teorema Bayes memiliki bentuk umum sebagai berikut:

$$P(H)|X = \frac{P(X) \mid H)P(H)}{P(X)} \quad \ldots (1)$$

X = Data dengan *class* yang belum diketahui
H = Hipotesis data X merupakan suatu *class* spesifik
P(H|X) = Probabilitas hipotesis H berdasarkan kondisi x (posteriori prob.)
P(H) = Probabilitas hipotesis H (prior prob.)
P(X|H) = Probabilitas X berdasarkan kondisi tersebut
P(X) = Probabilitas dari X

B. Algoritma *K-Nearest Neighbor* (KNN)

Algoritma *K-Nearest Neighbor* (K-NN) adalah suatu metode yang menggunakan algoritma *supervised*[5]. K-NN termasuk kelompok *instance-based learning.* Algoritma ini juga merupakan salah satu teknik *lazy learning.* K-NN dilakukan dengan mencari kelompok k objek dalam data *training* yang paling dekat (mirip) dengan objek pada data baru atau data *testing*[7].

C. Algoritma C4.5

Tree atau pohon banyak dikenal sebagai bagian dari *Graph*, yang termasuk dalam irisan bidang ilmu otomata dan teori bahasa serta matematika diskrit. *Tree* sendiri merupakan *graf* tak-berarah yang terhubung, serta tidak mengandung sirkuit.[8] Dalam sebuah *tree*, setiap pasang simpul terhubung hanya oleh satu lintasan, dan sebuah *tree* terdiri dari[9]:

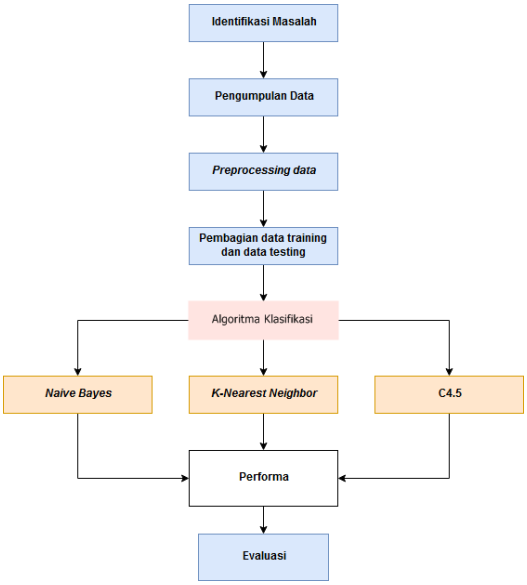
1. *Root*/akar, yang merupakan simpul tertinggi.
2. *Leaf*/daun, yang berupa simpul tanpa anak lagi
3. *Branch*/cabang, yang merupakan simpul-simpul selain daun.

Decision tree merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode *decision tree* mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami.

Decision tree juga berguna dalam mengekslorasi data, menemukan hubungan tersembunyi antara sejumlah calon *variable input* dengan sebuah variabel target. Karena *decision tree* memadukan antara eksplorasi data dan pemodelan. *Decision tree* digunakan untuk kasus-kasus dimana outputnya bernilai diskrit[6].

II. METODE PENELITIAN

Penelitian ini dilaksanakan di Perguruan Tinggi STMIK AMIKOM Purwokerto, data belum banyak diolah sehingga informasi terbaru masih sedikit diinstasi tersebut. Metode penelitian menggunakan metode komperatif, dimana penelitian ini bersifat membandingkan, dan pada penelitian ini dilakukan untuk membandingkan persamaan dan perbedaan dua atau lebih fakta-fakta yang diteliti berdasarkan kerangka pemikiran tertentu.



Gambar 2. Metode Penelitian.

Data penelitian ini diambil dari data kelulusan mahasiswa dari angkatan 2010-2012 yang diperoleh dari BAAK STMIK AMIKOM Purwokerto, sedangkan data nilai mahasiswa diambil dari semester 1-6 yang diperoleh dari bagian IT STMIK AMIKOM Purwokerto. Dan dalam penelitian ini menggunakan 2 dataset yang terdiri dari data 2.189 objek data dan 864 objek data, dimana data tersebut berisi 9 atribut kondisi dan 1 atribut keputusan.

III. HASIL DAN PEMBAHASAN

A. *Preprocessing Data*

Data-data yang diperoleh peneliti merupakan data-data yang masih terpisah sehingga perlu dilakukan penggabungan data agar menjadi satu file sehingga lebih mudah untuk dilakukan pengolahan data. Dalam proses pemilihan data atribut disesuaikan dengan apa yang sudah ditentukan pada bab sebelumnya.

Tahap pengolahan data yang pertama adalah pembersihan dan integrasi data. Proses pembersihan data dilakukan agar data-data yang diperoleh benar-benar data yang relevan sesuai dengan kebutuhan dan karena tidak semua atribut dalam tabel akan digunakan. Pembersihan data penting dilakukan untuk meningkatkan performa dalam proses *mining*. Cara pembersihan data dapat dilakukan dengan menghapus data-data yang tidak lengkap isiannya dan menghapus atribut-atribut yang tidak terpakai. Tabel dibawah ini bagian data kelulusan mahasiswa yang digunakan sebagai *dataset*.

Tabel 1. Atribut Kelulusan Mahasiswa dalam *dataset*

Atribut	Keterangan	Nilai Atribut
Jenis Kelamin	Menjelaskan jenis kelamin mahasiswa	Laki – laki, perempuan
Jurusan Sekolah	Menjelaskan jurusan mahasiswa saat di sekolah sebelumnya	IPA, IPS, Teknik, TIK, Lainnya
Program Studi	Menjelaskan pilihan program studi yang dipilih	Teknik Informatika, Sistem Informatika
IPS 1	Berisikan nilai mahasiswa dari semester 1	<2,5; 2,5 – 3,5; >3,5
IPS 2	Berisikan nilai mahasiswa dari semester 2	<2,5; 2,5 – 3,5; >3,5
IPS 3	Berisikan nilai mahasiswa dari semester 3	<2,5; 2,5 – 3,5; >3,5
IPS 4	Berisikan nilai mahasiswa dari semester 4	<2,5; 2,5 – 3,5; >3,5
IPS 5	Berisikan nilai mahasiswa dari semester 5	<2,5; 2,5 – 3,5; >3,5
IPS 6	Berisikan nilai mahasiswa dari semester 6	<2,5; 2,5 – 3,5; >3,5
Wisuda	Menjelaskan status mahasiswa apakah lulus tepat waktu atau tidak tepat waktu	Tepat Waktu, Tidak Tepat Waktu

Tabel 1 di atas merupakan tabel data kelulusan mahasiswa STMIK AMIKOM Purwokerto tahun angkatan 2010 hingga 2012. Data kelulusan mahasiswa tersebut diperoleh dari BAAK STMIK AMIKOM Purwokerto. Jumlah total mahasiswa dari tahun 2010 hingga 2012 yang sudah wisuda sebanyak 1.044 mahasiswa.

Data kelulusan mahasiswa tersebut tidak lantas langsung dapat dibangun menjadi sebuah *dataset* yang digunakan untuk memprediksi kelulusan mahasiswa. Diperlukan data-data penunjang lainnya yang menjadi faktor-faktor yang mempengaruhi tingkat kelulusan mahasiswa sehingga ada yang lulus tepat waktu dan ada yang lulus tidak tepat waktu. Data-data penunjang lainnya diantaranya ada data nilai mahasiswa dan data mahasiswa tahun angkatan 2010 hingga 2012 seperti yang ditunjukkan pada tabel-tabel di bawah ini.

Tabel 2. Data Nilai Mahasiswa Tahun 2010 – 2012

NIM	IPS1	IPS2	IPS3	IPS4	IPS5	IPS6
10.11.1466	-	2,75	3,18	3,00	2,55	3,00
10.11.1467	2,83	3,67	2,88	3,29	-	-
10.11.1468	-	2,50	2,91	2,27	2,36	2,09
10.11.1469	2,92	3,67	3,27	3,18	3,27	3,40
10.11.1471	3,33	3,17	3,55	3,45	3,45	3,70
10.11.1475	3,33	3,33	2,91	2,91	2,75	3,89
…	…	…	…	…	…	…
…	…	…	…	…	…	…
12.12.0367	1,96	-	1,41	0,29	-	-
12.12.0369	3,42	3,68	3,86	3,71	3,15	4,00
12.12.0370	2,75	2,32	2,68	2,14	1,00	2,52
12.12.0371	2,75	2,41	2,41	3,05	1,95	3,65
12.12.0372	0,33	-	2,55	0,86	-	-
12.12.2331	1,95	-	2,68	1,00	1,90	-

Nilai-nilai dalam *dataset* yang sudah diinput sebelumnya disesuaikan dengan format pemodelan *dataset* agar sesuai dengan pemodelan klasiifikasi. Setelah melewati beberapa tahap pengolahan data, dihasilkan *dataset* yang sudah tidak memiliki *missing value* maupun tipe data yang masih inkonsisten.

Setelah dilakukan proses transformasi data, langkah terakhir dari *preprocessing* data adalah mengubah *dataset* dari *file* excel menjadi format CSV atau ARFF agar dapat dikenali sebagai sumber data pada WEKA. Namun sebelum disimpan menjadi format file ARFF, diketahui bahwa

dataset yang ada masih merupakan *dataset* asli yang masih tercampur sehingga perlu dilakukan pembagian *dataset* menjadi 2 yaitu data yang akan digunakan sebagai data sampel (*data training*) dan data yang akan digunakan sebagai data uji/prediksi (*data testing*).

Bentuk *dataset* setelah penyesuaian atribut melalui tahap transformasi data dapat dilihat seperti pada tabel di bawah ini:

Tabel 3. *Dataset* Setelah Penyesuaian Atribut

JK	Jurusan Sekolah	Prodi	IP 1	IP 2	IP 3	IP 4	IP 5	IP 6	Wisuda
L	Teknik	TI	2	3	2	2	2	2	Tepat Waktu
L	Teknik	TI	2	2	3	2	2	3	Tepat Waktu
L	IPA	TI	2	2	2	2	2	3	Tepat Waktu
L	Teknik	TI	2	2	2	2	2	2	Tidak
L	Teknik	TI	3	2	2	2	2	2	Tidak
P	TIK	TI	3	3	3	3	2	2	Tidak
…	…	…	…	…	…	…	…	…	…
L	IPS	SI	1	2	2	2	2	2	Tepat Waktu
P	IPS	SI	2	2	2	2	2	2	Tidak
P	IPS	SI	2	2	2	2	2	3	Tepat Waktu
L	TIK	SI	2	2	2	2	2	2	Tidak
P	IPS	SI	3	3	2	2	2	2	Tidak
P	IPS	SI	2	3	3	2	2	2	Tepat Waktu

B. ***Pembagian Dataset***

Dataset yang sudah dibuat setelah melalui beberapa tahap *preprocessing* data terdiri dari 9 atribut kondisi dan 1 atribut kelas dengan jumlah *record* sebanyak 1.746 *record*. Dari 1.746 *record* ini, kelas label dengan nilai tepat waktu dan tidak tepat waktu berjumlah 667 *record* yang terdiri dari mahasiswa tahun angkatan 2010 dan 2011, dan mahasiswa yang belum lulus berjumlah 867 *record* yang terdiri dari mahasiswa tahun angkatan 2010 sampai dengan 2012. Karena tahun 2010 dan 2011 sudah melewati semester yang sudah dijadwalkan yaitu 8 semester sehingga mahasiswa yang belum lulus sudah pasti lulus tidak tepat waktu. Sedangkan mahasiwa yang belum lulus tahun dari tahun 2010 hingga 2012 akan digunakan sebagai *data testing* untuk memprediksi apakah mahasiswa tersebut akan lulus tepat waktu atau tidak tepat waktu. Sehingga dalam penelitian ini *dataset* akan dibagi menjadi 2, yaitu data yang akan digunakan sebagai data sampel (*data training*) dan data yang akan digunakan sebagai data uji/prediksi (*data testing*).

a. *Data Training*

Data training merupakan *dataset* yang berisi 667 *record* yang terdiri dari *dataset* kelulusan mahasiswa tahun angkatan 2010 – 2011 baik yang lulus tepat waktu maupun yang lulus tidak tepat waktu.

Data training yang sudah dibuat disimpan dalam format file ARFF. *Data training* berisi 667 *record* yang berasal dari *dataset* kelulusan tahun angkatan 2010 – 2011 yang sudah lulus baik yang lulus tepat waktu maupun yang lulus tidak tepat waktu. Dari 667 *record*, 332 *record* berasal dari angkatan 2010 dan 335 *record* dari angkatan 2011.

b. *Data Testing*

Data testing merupakan *dataset* yang berisi 867 *record* yang terdiri dari *dataset* mahasiswa tahun angkatan 2010 – 2012 yang belum lulus. Selanjutnya *data testing* yang sudah dibuat. Berisi 867 *record* data yang merupakan *dataset* kelulusan mahasiswa tahun angkatan 2010 hingga 2012 yang belum lulus, sehingga data tersebut dijadikan sebagai *data testing* untuk memprediksi kelulusan mahasiswa apakah tepat waktu atau tidak.

C. Penerapan Algoritma Klasifikasi

Setelah dilakukan pembagian *dataset* akan dilakukan proses *mining* atau pengujian *dataset* menggunakan metode atau algoritma yang sudah ditentukan. Proses *mining* bertujuan untuk mendapatkan pola-pola dan informasi yang tersembunyi di dalam basis data yang telah melewati tahap *preprocessing data*.

Proses *mining* dilakukan dengan menerapkan metode *Naive Bayes*, K-*Nearest Neighbor* dan C4.5 serta menggunakan sistem cerdas yang disediakan oleh *tool* WEKA. Pengujian akan dilakukan terhadap masing-masing *dataset* yang sudah dibagi pada pembahasan sebelumnya.

D. *Uji Akurasi*

a. Pengujian *Data training*

Data training yang merupakan data yang terdiri dari data kelulusan mahasiswa dan data nilai dari tahun 2010 – 2011 dengan jumlah total 667 *record*. Data tersebut akan diolah dengan *tool* WEKA menggunakan algoritma *Naive Bayes, K-Nearest Neighbor* dan C4. Pengujian yang sudah dilakukan terhadap *data training* menghasilkan akurasi tertinggi sebesar 88,16% dari metode *k-nearest neighbor,* sedang algoritma *naive bayes* memiliki prediksi 79,46% dan algoritma c4.5 sendiri memprediksi dengan nilai akurasi sebesar 82,31%, hasil pengujian terhadap *data training* menggunakan WEKA dapat dilihat pada tabel di bawah ini.

b. Pengujian *data Testing*

Setelah dilakukan pengujian terhadap *data training* kemudian dilanjutkan pengujian terhadap *data testing* untuk memprediksi tingkat kelulusan mahasiswa apakah tepat waktu atau tidak.

Data Testing yang merupakan data yang terdiri dari data kelulusan mahasiswa dan data nilai dari tahun 2010 – 2011 dengan jumlah total 867 *record*. Data tersebut akan diolah dengan *tool* WEKA menggunakan algoritma *Naive Bayes, K-Nearest Neighbor* dan C4. Pengujian yang sudah dilakukan terhadap *data testing* menghasilkan akurasi tertinggi sebesar 89,04% dari metode *k-nearest neighbor,* dan dari hasil perhitungan manual yang sudah peneliti lakukan dihasilkan data dengan label "Tepat Waktu" sebanyak 772 *record* sedangkan dengan label "Tidak" berjumlah 95 *record*. Hal ini berarti akurasi yang dihasilkan dapat dihitung sebagai berikut:

Prosentasi akurasi = (Total prediksi benar / Total data) x 100%
= (772 / 867) x 100%
= 89,04%

Sedangkan hasil perhitungan yang dilakukan menggunakan Weka dihasilkan akurasi dengan prosentase 89,04% dimana dari total 867 *record* diperoleh 772 *record*

yang memiliki label "Tepat Waktu" sedangkan 95 *record* yang memiliki label "Tidak". Maka dapat diambil kesimpulan bahwa perhitungan manual yang dilakukan oleh peneliti dengan perhitungan yang dilakukan menggunakan Weka memperoleh hasil yang sama.

E. Evaluasi Hasil Akurasi

Tabel 3.5, merupakan tabel hasil pengujian terhadap *data testing* dari masing-masing algoritma yang telah diuji dengan perhitungan *confusion matrix* sebagai berikut:

a. Berdasarkan tabel tersebut bahwa hasil evaluasi tingkat akurasi algoritma *Naive Bayes* sebesar 79,82%, dan tingkat akurasi tersebut diperoleh dari hasil perhitungan sebagai berikut:

Akurasi = (TP + TN) / (TP + FP + FN + TN)
= (645 + 47) / (645 +71 + 104 + 47)
= 692 / 867
= 79,82%

b. Sedangkan pengujian dari hasil evaluasi tingkat akurasi algoritma *K-Nearest Neighbor* sebesar 89,04%, dan tingkat akurasi tersebut diperoleh dari hasil perhitungan sebagai berikut:

Akurasi = (TP + TN) / (TP + FP + FN + TN)
= (730 + 42) / (730 + 76 + 19 + 42)
= 772 / 867
= 89,04%

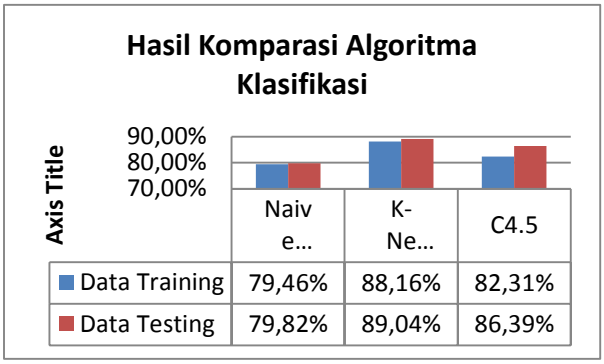
c. Dan hasil evaluasi tingkat akurasi algoritma C4.5 sebesar 79,82%, dan tingkat akurasi tersebut diperoleh dari hasil perhitungan sebagai berikut:

Akurasi = (TP + TN) / (TP + FP + FN + TN)
= (749 + 0) / (749 + 118 + 0 + 0)
= 749 / 867
= 86,39%

F. Hasil Analisa

Penelitian yang sudah dilakukan dengan menggunakan *dataset* kelulusan dan data nilai yang terdiri dari *data training* dan *data testing*, dengan menerapkan metode k-optimal pada 2 jenis *dataset* tersebut menghasilkan perolehan nilai akurasi pada *data training* yaitu sebesar 88,16%, sedangkan pada pengujian *data testing* menghasilkan akurasi yaitu sebesar 89,04% dari metode *K-Nearest Neighbor*.

Kemudian jika dilihat dari nilai yang dihasilkan oleh *data testing* pada penelitian yang dilakukan oleh Peneliti menggunakan metode *Naive Bayes*, K-*Nearest Neighbor* dan C4.5 menghasilkan akurasi yang lebih besar yaitu dari algoritma *K-Nearest Neigbor* dengan prediksi 89,04% sedangkan menggunakan metode klasifikasi *Naive Bayes* dan C4.5 dari *data testing* hanya menghasilkan akurasi sebesar 79,82%. Dan 86,39%. Perbandingan hasil akurasi juga dapat dilihat melalui grafik yang ditunjukkan pada gambar 4.1 dibawah ini.



Gambar 3. Hasil Komparasi Algoritma Klasifikasi

V KESIMPULAN

Dari total data asli yang sudah digabungkan berjumlah 2.189 *record*, kemudian dilakukan pembersihan dan integrasi data diperoleh data sebanyak 1.736 *record*. Dari 1.736 *record* dibagi menjadi 2 dataset yaitu *data training* dan *data testing*, data training berjumlah 667 *record* yang berasal dari mahasiswa tahun 2011 - 2011 yang sudah lulus baik tepat waktu maupun tidak tepat waktu dan *data testing* berjumlah 867 *record* dari mahasiswa 2010 - 2012 yang belum lulus. Pengujian yang sudah dilakukan terhadap *data training* dengan total data 667 menghasilkan akurasi tertinggi sebesar 88,16% dari metode *k-nearest neighbor*, sedang algoritma *naive bayes* memiliki prediksi 79,46% dan algoritma c4.5 sendiri memprediksi dengan nilai akurasi sebesar 82,31%, hasil pengujian terhadap *data training* menggunakan WEKA Dari hasil penerapan algoritma klasifikasi *data mining* dengan pengujian jenis *data testing* dari total data 867 *record*, perhitungan *confusion matrix* menunjukkan bahwa algoritma *K-Nearest Neighbor* memiliki tingkat akurasi yang lebih baik dengan prediksi akurasi sebesar 89,04%, dibandingkan dengan algoritma *naive bayes* hanya memiliki tingkat prediksi 79,82%, dan algoritma C4.5 memiliki prediksi sekitar 86,39%.

DAFTAR PUSTAKA

[1]. STMIK AMIKOM Purwokerto, Buku Panduan Akademik Mahasiswa Tahun Ajaran 2012-2013. Purwokerto, Jawa Tengah: STMIK AMIKOM Purwokerto, 2012.

[2]. Nugroho. 2014. *Penerapaan Algoritma C4.5 Untuk Klasifikasi Predikat Kelulusan Mahasiswa fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta*.Prosiding Seminar Nasional Sain & Teknologi (SNAST), ISNN:1979-911X.

[3]. Fiastantyo, Giat. 2014. *Perbandingan Kinerja Metode Klasifikasi Data Mining Menggunakan Naive Bayes dan Algoritma C4.5 Untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa*. Universitas Dian Nuswantoro.

[4]. Hastuti, Khafiizh. 2012. Analisis Komparasi Algoritma Klasifikasi *Data Mining* Untuk Prediksi Mahasiswa Non Aktif. *ISBN 979-26-0255-0*.

[5]. Han, J., & Kamber, M. 2006. Data Mining Concept and Tehniques San Fransisco: Morgan Kauffman. ISBN 13: 978-1-55860-901-3

[6]. Kusrini, Luthfi, E.T. (2009). "*Algoritma Data Mining"*, Andi Offset. Surabaya.

[7]. Leidiyana. Penerapan Algoritma *K-Nearest Neighb*or Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bemotor. Jurnal Penelitian Ilmu Komputer, System Embedded & Logic, Vol : 1. STMIK Nusa Mandiri. 2010

[8]. Munir, R (2010). *Matematika Diskrit.* Bandung: Informatika Bandung.

[9] Utdirartatmo, F (2005). *Teori Bahasa dan Otomata*, Yogyakarta: Graha Ilmu.