Klasifikasi Lama Masa Studi Mahasiswa Menggunakan Perbandingan Metode Algoritma C.45 dan Algoritma *Classification and Regression Tree*

Comparison of C4.5 Algorithm and Classification and Regression Tree Algorithm In The Classification of Study Period of Undergraduate Students

Hadi Dwi Darmawan[1], Desi Yuniarti[2], dan Yuki Novia Nasution[3]
[1]Laboratorium Statistika Komputasi FMIPA Universitas Mulaw*arm*an
[2,3]Jurusan Matematika FMIPA Universitas Mulaw*arm*an
[1]E – mail: hadidwid*arm*awanstatistika13@gmail.com

Abstract

Classification is the grouping samples based on the characteristics of the similarities and differences using target variable category. In this study, the decision tree is formed using C4.5 algorithm and Classification and regression tree (CART) algorithm to classify a student's study period class of 2016 FMIPA UNMUL. C4.5 algorithm is a non binary classification tree where the branches of trees can be more than two on C4.5 algorithm, decision tree is established based on Entropy value. The purpose of CART algorithm is to get an accurate data as group identifier of a classification. CART can be applied in three main steps, namely the establishment of a classification tree, trimming of the classification tree, and determination of optimal classification tree. The main goal of this research is to determine factors which may effect on all predicate graduation who was graduated on 2016 using C4.5 algorithm and CART algorithm and also to know comparison accuracy of classification result by C4.5 algorithm and CART algorithm. The result showed that factors which affected the duration of all graduation using C4.5 algorithm are major (X4), region school (X5) and region origin (X3) and factors affected to the duration of all graduation using CART algorithm are major (X4) and Cumulative Achievement Index (X1). Precision classification in CART algorithm is better than C4.5 algorithm. C4.5 algorithm was able to predict with 40% accuracy while the CART algorithm has a predictive accuracy of 60%.

Keywords: C4.5 algorithm, CART algorithm, classification, study periode, decision tree

Pendahuluan

Data adalah catatan atas kumpulan fakta. Dalam penggunaan sehari-hari data berarti suatu pernyataan yang dapat diterima. Data dapat diperoleh, disimpan, diolah, dipakai dan seba*gain*ya. Salah satu bentuk pengolahan suatu data yaitu data mining. Data mining suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. Data mining juga merupakan proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk menguraikan, mengidentifikasi informasi yang bermanfaat, dan pengetahuan yang terkait dari berbagai *database* besar (Efraim, 2005).

Decision tree (pohon keputusan) adalah pohon klasifikasi yang digunakan sebagai prosedur penalaran untuk mendapatkan jawaban dari masalah yang dikasuskan dalam *decision tree*, daerah pengambilan keputusan yang sebelumnya kompleks dapat diubah menjadi lebih sederhana. Banyak algoritma yang dapat dipakai dalam pembentuknya *decision tree* seperti ID3, CART, C4.5, dan lain-lain. Algoritma adalah urutan langkah-langkah yang logis untuk menyelesaikan suatu masalah (Prasetyo, 2012).

Algoritma CART adalah salah satu metode atau Algoritma dari salah satu teknik eksplorasi data yaitu teknik pohon keputusan Algoritma CART dikembangkan untuk melakukan analisis klasifikasi pada variabel respon baik yang nominal, ordinal maupun kontinu. Algoritma CART juga dapat menyeleksi variabel dan interaksi-interaksi variabel yang paling penting dalam menentukan hasil atau variabel prediktor (Breiman et al, 1993). Model yang dihasilkan cukup sederhana untuk menerangkan suatu amatan dikelompokkan atau diduga dalam kelompok tertentu (Statsoft, 2003). Sedangkan menurut Yohannes dan Hoddinott (1999), kelemahan dari Algoritma CART adalah hasil akhirnya tidak didasarkan model probabilistik. Tidak ada tingkat probabilitas dari selang kepercayaan yang berhubungan dengan dugaan yang didapat dari pohon Algoritma CART untuk pengelompokkan data baru. (Yohannes dan Hoddinott, 1999).

Penelitian terdahulu mengenai metode Algoritma CART dan Algoritma C4.5 telah dilakukan antara lain oleh Siahaan (2015) tentang Aplikasi *Classification and regression tree* dan Regresi Logistik Ordinal dalam Bidang Pendidikan dengan studi kasus Predikat Kelulusan Mahasiswa S1 Fakultas Matematika dan Ilmu pengetahuan Alam Universitas Mulaw*arm*an dan Chair (2016) tentang Aplikasi Klasifikasi Algoritma C4.5 dengan studi kasus Masa Studi Mahasiswa Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Mulaw*arm*an Angkatan 2008. Perbedaan penelitian dengan penelitian

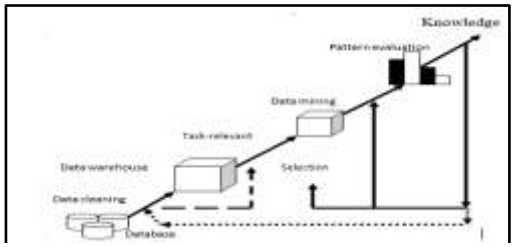
sebelumnya yaitu terletak pada perbandingan metode yang digunakan.

Data Mining

Data mining adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Larose, 2005).
Menurut Tan (2006), secara garis besar data mining dapat dikelompokkan menjadi 2 kategori utama, yaitu:

1. *Descriptive mining*, yaitu proses untuk menemukan karakteristik penting dari data dalam suatu basis data. Teknik data mining yang termasuk dalam descriptive mining adalah clustering, association, dan sequential mining.
2. *Predictive*, yaitu proses untuk menemukan pola dari data dengan menggunakan beberapa variabel lain di masa depan. Salah satu teknik yang terdapat dalam predictive mining adalah klasifikasi.

Secara sederhana data mining bisa dikatakan sebagai proses menyaring atau "menambang" pengetahuan dari sejumlah data yang besar. Istilah lain untuk data mining adalah *Knowledge Discovery in Database* atau KDD. Walaupun sebenarnya data mining sendiri adalah bagian dari tahapan proses dalam KDD, seperti yang terlihat pada Gambar 1 (Han dan Kamber, 2001).



Gambar 1. Data mining sebagai salah satu tahapan dalam proses *Knowledge Discovery in Database*

Menurut Thomas (2004), tujuan dari adanya data mining adalah :

1. Explanatory, yaitu untuk menjelaskan beberapa kegiatan observasi atau suatu kondisi.
2. Confirmatory, yaitu untuk mengkonfirmasi suatu hipotesis yang telah ada.
3. Exploratory, yaitu untuk menganalisis data baru suatu relasi yang janggal.

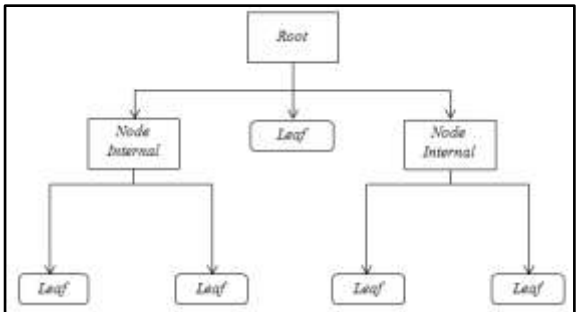
Menurut Larose (2005), data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu deskripsi, estimasi, prediksi, klaster, asosiasi dan klasifikasi. (Larose, 2005).

Decision tree

Classification *decision tree* merupakan teknik klasifikasi yang sederhana dan banyak digunakan. Bagian ini membahas bagaimana *decision tree* bekerja dan bagaimana *decision tree* dibangun. Seringkali untuk mengklasifikasikan obyek, diajukan urutan pertanyaan sebelum bisa ditentukan kelompoknya. Jawaban pertanyaan pertama akan mempengaruhi pertanyaan berikutnya dan seterusnya. Dalam *decision tree*, pertanyaan pertama akan ditanyakan pada simpul akar pada level 0. Jawaban dari pertanyaan ini dikemukakan dalam cabang-cabang. Jawaban dalam cabang akan disusul dengan pertanyaan kedua lewat simpul yang berikutnya pada level 1. Dengan memperhatikan *decision tree* dalam Gambar 2 akan nampak ada 4 level pertanyaan. Dalam setiap level ditanyakan nilai atribut melalui sebuah simpul. Jawaban dari pertanyaan itu dikemukakan lewat cabang-cabang. Langkah ini akan berakhir di suatu simpul jika pada simpul tersebut sudah ditemukan kelas atau jenis obyeknya. Jika dalam satu tingkat suatu obyek sudah diketahui termasuk dalam kelas tertentu, maka proses berhenti di level tersebut. Jika tidak, maka dilanjutkan dengan pertanyaan di level berikutnya hingga jelas ciri-cirinya dan jenis obyek dapat ditentukan (Santosa, 2007).

Karakteristik dari *decision tree* dibentuk sejumlah elemen sebagai berikut (Tan, 2006).

1. *Node*, yang menyatakan variabel, *Node* bisa berupa variabel akar, variabel cabang, dan kelas.
2. *Arm*, setiap cabang menyatakan nilai hasil pengujian di *node* bukan *leaf*.
3. *NodeRoot*, tidak mempunyai input *arm* yaitu lengan masukan dan mempunyai nol atau lebih output *arm* yaitu lengan keluar.
4. *Nodeinternal*, setiap *node* yang bukan *leaf* (non terminal) yang mempunyai tepat satu input *arm* dan dua atau lebih output *arm*, *node* ini menyatakan pengujian yang didasarkan pada nilai fitur.
5. *Nodeleaf* (terminal) adalah *node* yang mempunyai tepat satu input *arm* dan tidak mempunyai output *arm*. *Node* ini menyatakan label kelas (keputusan).



Gambar 2.Contoh karateristik *decision tree*

Algoritma C4.5

Algoritma merupakan kumpulan perintah yang tertulis secara sistematis guna menyelesaikan permasalahan logika dari matematika. Pengertian Algoritma C4.5 adalah algoritma yang digunakan untuk membentuk *decision tree*. *Decision tree* dapat diartikan suatu cara untuk memprediksi atau mengklasifikasi yang sangat kuat. *Decision tree* dapat membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapkan serangkaian aturan keputusan (Fayyad, 1996).

Secara umum Algoritma C4.5 untuk membangun *decision tree* adalah sebagai berikut (Kusrini dan Luthfy, 2009) :

1. Pemilihan variabel akar, Untuk memilih variabel akar, didasarkan pada nilai *gain* tertinggi dari variabel-variabel yang ada. Berikut adalah cara untuk menghitung nilai *gain*:

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

   Sebelum mendapatkan nilai *Gain*, dicari terlebih dahulu nilai Entropy. Entropy adalah informasi mengenai proporsi pembagian kelas, nilai Entropy berkisar mulai dari 0 sampai dengan 1, jika nilai Entropy = 0, maka menandakan jumlah sampel hanya berada di salah satu kelas, sedangkan jika nilai Entropy = 1, maka menandakan jumlah sampel berada di masing-masing kelas dengan jumlah yang sama. Adapun rumus dasar dari perhitungan Entropy adalah sebagai berikut :

   $$Entropy(S) = -\sum_{i=1}^{n} P_i \quad {}^{2}Log\ P_i \quad (2)$$
2. Penentuan cabang untuk masing-masing nilai, Untuk penentuan cabang sama seperti mencari variabel akar yaitu didasarkan pada nilai *gain* tertinggi dari variabel-variabel yang ada.
3. Kelas dibagi dalam cabang dan apabila cabang mempunyai dua kelas maka dipilih kelas yang terbanyak
4. Proses diulang untuk masing-masing cabang sampai mana semua kelas pada cabang memiliki kelasnya masing-masing

Classification and regression tree

Classification and regression tree (CART) adalah salah satu metode atau algoritma dari salah satu teknik eksplorasi data yaitu teknik *decision tree*. CART dikembangkan untuk melakukan analisis klasifikasi pada variabel respon baik yang nominal, ordinal, maupun kontinu. CART menghasilkan suatu pohon klasifikasi jika variabel responnya kategorik dan menghasilkan pohon regresi jika variabel responnya kontinu. Prinsip dari metode pohon klasifikasi ini adalah memilah seluruh amatan menjadi dua gugus amatan dan memilah kembali gugus amatan tersebut menjadi dua gugus amatan berikutnya, hingga diperoleh jumlah amatan minimum pada tiap-tiap gugus amatan berikutnya. Tujuan utama CART ialah untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian (Timofeev, 2004).

CART mempunyai beberapa kelebihan dibandingkan metode klasifikasi lainnya, yaitu hasilnya lebih mudah diintreprestasikan, lebih akurat dan lebih cepat perhitungannya, selain itu CART bisa diterapkan untuk himpunan data yang mempunyai jumlah besar, variabel yang sangat banyak dengan skala variabel campuran melalui prosedur pemilahan biner (Lewis, 2000).. Metode CART menurut Lewis (2000) memiliki kelemahan sebagai berikut :

1. CART mungkin tidak stabil dalam *decision tree* karena CART sangat sensitif dengan data baru.
2. CART sangat bergantung dengan jumlah sampel. Jika sampel data learning dan data testing berubah maka *decision tree* yang dihasilkan juga ikut berubah.
3. Tiap pemilahan bergantung pada nilai yang hanya berasal dari satu variabel penjelas.

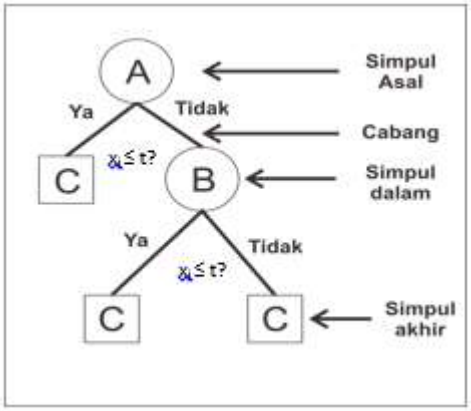
Decision tree dibentuk dengan menggunakan algoritma penyekatan rekursif (berulang) dan secara binary recursive partitioning (biner). Pemilahan dilakukan untuk memilah data menjadi dua kelompok, yaitu kelompok yang masuk simpul kiri dan yang masuk simpul kanan. Pemilahan dilakukan pada tiap simpul sampai didapatkan suatu simpul terminal atau akhir. Variabel yang memilah pada simpul utama adalah variabel terpenting dalam menduga kelas amatan (Lewis and Roger, 2000).

Pembentukan Pohon Klasifikasi

Proses pembentukan pohon klasifikasi terdiri atas tiga tahapan, yaitu(Lewis and Roger, 2000) :

a. Pemilihan (*Classifier*)
b. Untuk membentuk pohon klasifikasi digunakan sampel data *learning* (*L*) yang masih bersifat heterogen. Setiap pemilihan hanya bergantung pada nilai yang berasal dari suatu variabel independen. Struktur pohon klasifikasi dapat dilihat pada Gambar 3. Penentuan *Node* Terminal

   Suatu *node* t akan menjadi *node* terminal atau tidak, akan dipilah kembali bila pada *node* t tidak terdapat penurunan keheterogenan secara berarti atau adanya batasan minimum n seperti hanya terdapat satu pengamatan pada tiap *node* anak. Umumnya jumlah kasus minimum dalam suatu terminal akhir adalah 5. Dan apabila hal itu terpenuhi maka pengembangan pohon dihentikan.



Gambar 3. *Data mining* sebagai salah satu tahapan dalam proses *Knowledge Discover*

c. Penandaan Label Kelas

Penandaan label *node* terminal berdasar aturan jumlah anggota kelas terbanyak, yaitu :

$$p(j_0|t) = max_j\ p(j|t) = max_j \frac{N_j(t)}{N(t)} \quad (3)$$

dengan $p(j|t)$ adalah proporsi kelas j pada *node t*, $N_j(t)$ adalah jumlah pengamatan kelas j pada *node* t dan $N(t)$ adalah jumlah pengamatan pada *node*t. Label kelas *node* terminal t adalah j_0 yang memberi nilai dugaan kesalahan pengklasifikasian *node t* terbesar.

Proses pembentukan pohon klasifikasi berhenti saat terdapat hanya satu pengamatan dalam tiap simpul anak atau adanya batasan minimum n, semua pengamatan dalam tiap *node* anak identik, dan adanya batasan jumlah kedalaman pohon maksimal. Setelah terbentuk pohon maksimal tahap selanjutnya adalah pemangkasan pohon untuk mencegahnya terbentuk pohon klasifikasi yang berukuran sangat besar dan kompleks, sehingga diperoleh ukuran pohon yang layak.(Lewis and Roger, 2000).

Indeks *Gini*

Aturan pemisahan Indeks Gini adalah aturan yang paling luas dan yang paling sering digunakan. Aturan ini mengikuti impurity i(t), dimana fungsinya adalah suatu fungsi yang digunakan untuk mengukur keakuratan model dengan memberikan indikasi kehomogenan kelas-kelas pada data sehingga pada terminal *node* akhir akan didapatkan data yang lebih akurat dalam sebuah *node* t, misal terdapat 1,2…, j kelas.

Sebagai penggunaan aturan pluralitas untuk mengklasifikasikan objek dalam *tnode*, menggunakan aturan pluralitas untuk mengklasifikasikan objek dalam *t node*, menggunakan aturan yang memberikan objek yang dipilih secara acak dari *node* atau simpul ke kelas *j* dengan probabilitas adalah $p(j|t)$. Probabilitas memperkirakan bahwa item sebenarnya di kelas *i* adalah $p(i|t)$. Oleh karena itu, kemungkinan perkiraan kesalahan klasifikasi di bawah aturan ini adalah indeks *Gini* seperti pada persamaan (4).

Node t dibelah menjadi 2 subset D_1dan D_2 dengan ukuran masing-masing $n_1$dann_2, indeks *gini* dari pembelahan tersebut didefinisikan sebagai berikut :

$$Gini_{pembela\ han}(t) = \frac{n_1}{n} Gini(D_1) + \frac{n_2}{n} Gini(D_2) \quad (4)$$

(Breimen dkk, 1993).

Pemangkasan Pohon Klasifikasi

Bagian pohon yang kurang penting dilakukan pemangkasan sehingga didapatkan pohon klasifikasi yang optimal. Pemangkasan didasarkan pada suatu penilaian ukuran sebuah pohon tanpa mengorbankan ketepatan melalui pengurangan *node*pohon sehingga dicapai ukuran pohon yang layak. Ukuran pemangkasan yang digunakan untuk memperoleh ukuran pohon yang layak tersebut adalah *cost complexity minimum* (Breimen dkk, 1993).

Sebagai bayangan, untuk sembarang pohon *T* yang merupakan sub pohon dari pohon terbesar *Tmax* diperoleh ukuran *complexity*-nya yaitu $|\tilde{T}|$ yang menyatakan banyaknya *node* akhir pada pohon *T* tersebut. Dalam regresi biasa,$|\tilde{T}|$ analog dengan derajat bebas model. Untuk suatu $\alpha \geq 0, \alpha \epsilon R^1$ yang disebut sebagai *complexity parameter* yakni ukuran tentang *cost* bagi penambahan satu *node* akhir pada pohon *T*, maka ukuran *cost complexity*-nya adalah :

$$R_\alpha(t) = R(T) + \alpha|\tilde{T}| \quad (5)$$

dimana $R_\alpha(t)$ merupakan kombinasi linier nilai dan kompleksitas pohon yang dibentuk dengan menambahkan *cost penalty* bagi kompleksitas terhadap nilai bagi kesalahan klasifikasi pohon. *Cost penalty pruning* menentukan suatu pohon bagian *T*(α) yang meminimumkan $R_\alpha(t)$ pada seluruh bagian, atau untuk setiap nilai α, dicari pohon bagian *T*(α) <T_{max} yang meminimumkan $R_\alpha(t)$ yaitu :

$$R_\alpha(T(\alpha)) = min_{T<T_{max}} R_\alpha(T) \quad (6)$$

Jika *R*(*T*) digunakan sebagai kriteria penentuan pohon optimal maka akan cenderung pohon terbesar adalah T_1, sebab semakin besar pohon, maka semakin kecil nilai *R*(*T*). *Jumping-off point* bagi pemangkasan bukan T_{max} melainkan T_1 yakni suatu sub pohon yang memenuhi kriteria $R(T_1) = R(T_{max})$. untuk mendapatkan T_1 *dan* T_{max}, ambil t_L dan t_R yang merupakan *node* dari anak kiri dan *node* anak kanan dari T_{max}yang dihasilkan dari *node*induk *t*. Kemudian hitung nilai t_L dan t_R menggunakan persamaan :

$$R(t) = r(t)P(t) \quad (7)$$

Dengan nilai *r*(*t*) = 1 – max $_j$ $P(j|t)$ dan $P(j|t)$adalah peluang beberapa objek yang berada dalam *node*, begitu diperoleh dua *node* anak dan *node*induknya yang memenuhi persamaan :

$$R(t) = R(t_L) + R(t_R) \quad (8)$$

Maka pangkas *node* anak t_L dan t_R tersebut (Breiman dkk, 1993).

Penentuan Pohon Klasifikasi Optimal

Penentuan klasifikasi yang berukuran besar akan memberikan nilai penduga pengganti paling kecil, sehingga pohon ini cenderung dipilih untuk menduga nilai respon. Tetapi ukuran pohon yang besar akan menyebabkan nilai kompleksitas yang tinggi karena struktur data yang digambarkan cenderung kompleks, sehingga perlu dipilih pohon optimal yang berukuran sederhana tetapi memberikan nilai penduga pengganti cukup kecil. Ada dua jenis penduga pengganti yaitu penduga sampel uji independen dan penduga validasi silang lipat V (*cross validation V-fold estimate*)(Breiman dkk, 1993).

a. Cross Validation V-Fold Estimate

Penduga validasi silang lipat V sering digunakan apabila amatan yang tidak cukup besar. Amatan dalam *L* dibagi secara acak menjadi *V* bagian yang saling lepas dengan ukuran kurang lebih sama besar untuk setiap kelasnya. Pohon $T^{(V)}$ dibentuk dari $L - L_V$ dengan v = 1,2…,*V*. Misalkan $d^{(v)}(x)$ adalah hasil pengklasifikasian, penduga sampel uji untuk $R(T_1^{(v)})$ yaitu :

$$R^{ts}\left(T_t^{(v)}\right) = \frac{1}{Nv}\sum_{(x_n j_n)} X(d^{(v)}(x_n) \neq j_n) \quad (9)$$

dengan $Nv = \frac{N}{V}$adalah jumlah amatan dalam *Lv*. Kemudian dilakukan prosedur yang sama menggunakan seluruh *L*, maka penduga validasi silang lipat *V* untuk $T_t^{(v)}$ adalah :

$$R^{Cv}(T_t) = \frac{1}{V}\sum_{v=1}^{v} R^{ts}(T^{(v)}) \quad (10)$$

Pohon klasifikasi optimum dipilih T^* dengan,

$$R^{Cv}(T^*) = min_t\ R^{Cv}(T_t) \quad (11)$$

(Breiman dkk, 1993).

Ketepatan Klasifikasi

Apparent Error Rate (APER) atau yang disebut laju error merupakan ukuran evaluasi yang digunakan untuk melihat peluang kesalahan klasifikasi yang dihasilkan oleh suatu fungsi klasifikasi. Nilai APER menunjukkan proporsi observasi yang salah diklasifikasikan oleh fungsi klasifikasi. Semakin kecil nilai APER maka hasil pengklasifikasian semakin baik (Prasetyo, 2012).
Menurut Johnson dan Wichern (2007), terjadinya kesalahan klasifikasi suatu observasi merupakan hal yang sangat mungkin terjadi. Hal ini dikarenakan terkadang terdapat beberapa observasi yang tidak berasal dari kelompok tertentu tetapi dimasukkan ke dalam kelompok tersebut. Perhitungan nilai APER dapat dilakukan dengan menggunakan matriks konfusi sebagaimana Tabel 1.

Tabel 1. Matrix Confusion

Kelompok Aktual	Kelompok Prediksi		Jumlah Observasi
	1	2	
1	n_{11}	n_{12}	n_1
2	n_{21}	n_{22}	n_2

$$APER = \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} \quad (3)$$

n_{11}=banyak data dalam kelompok 1 yang secara benar dipetakan ke kelompok 1
n_{12} = banyak data dalam kelompok 1 yang secara salah dipetakan ke kelompok 2
n_{21} =banyak data dalam kelompok 2 yang secara salah dipetakan ke kelompok 1
n_{22}=banyak data dalam kelompok 2 yang secara benar dipetakan ke kelompok 2
dapat dilihat perhitungan nilai APER yang telah diuraikan tersebut, maka dapat dilihat nilai *error*-nya, sehingga untuk mencari nilai ketepatannya dapat menggunakan 1 - APER.

Lama Masa Studi

Pendidikan adalah suatu aktivitas sosial yang memungkinkan masyarakat berkembang. Pendidikan dapat didefinisikan sebagai keseluruhan pengalaman belajar setiap orang sepanjang hidupnya yang berlangsung tidak dalam batas usia tertentu tetapi berlangsung sepanjang hidup (Mudyaharjo, 2002). Pendidikan di Indonesia terdiri dari beberapa jenjang. Salah satu jenjang pendidikan yang menjadi syarat dasar dalam pekerjaan saat ini adalah perguruan tinggi, di mana perguruan tinggi akan mempersiapkan calon-calon sarjana yang handal dan mempunyai keterampilan di bidangnya(Mudyaharjo, 2002).

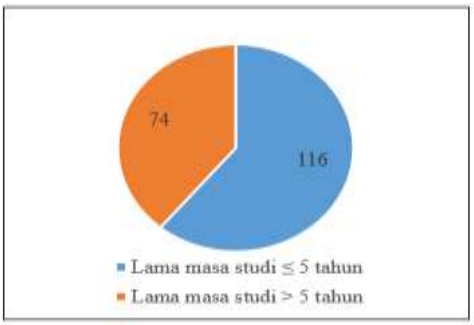
Menurut Kamus Besar Bahasa Indonesia (KBBI), istilah sarjana berasal dari Bahasa Sangkakerta, yang merupakan sebuah gelar akademik yang diberikan kepada lulusan program pendidikan strata satu atau undergraduate. Untuk program Strata Satu (S1), persyaratan penyelesaian studi ditetapkan sekurang-kurangnya telah menempuh 144 SKS dan sebanyak-banyaknya 160 SKS, dengan lama masa studi paling lama 7 tahun akademik untuk program sarjana dan program diploma 4 atau sarjana terapan (Standar Nasional Pendidikan Tinggi Nomor 44 Tahun 2015). Waktu yang dibutuhkan untuk memperoleh gelar sarjana ini sering disebut dengan Lama Masa Studi. Apabila terdapat seorang mahasiswa yang lama masa studinya lebih dari 7 tahun, maka mahasiswa tersebut akan mengalami Drop Out (DO).(Mudyaharjo, 2002)

Hasil dan Pembahasan

Analisis Deskriptif

Data yang digunakan dalam penelitian kali ini adalah data pada kelulusan seluruh mahasiswa FMIPA UNMUL tahun 2016. Data kelulusan pada tahun 2016 ada sebanyak 190 sampel yang diambil

dari akademik FMIPA UNMUL. Variabel respon adalah Lama Masa Studi (*Y*), dengan faktor yang diduga mempengaruhi Lama Masa Studi (*Y*) adalah Indeks Prestasi Kumulatif (IPK) (*X1*), jenis kelamin (*X2*), asal daerah (*X3*), Program Studi (*X4*), dan asal sekolah (*X5*).



Gambar 4. Diagram Lingkaran untuk Variabel Lama Masa Studi

Berdasarkan Gambar 4 dapat diketahui bahwa mahasiswa yang lulus dengan Lama Masa Studi kurang dari sama dengan 5 tahun ada 116 mahasiswa dan mahasiswa yang lulus dengan Lama Masa Studi lebih dari 5 tahun ada 74 mahasiswa dari total sebanyak 190 mahasiswa lulusan tahun 2016 FMIPA UNMUL.

Analisis deskriptif ini dilakukan untuk mengetahui karakteristik data Lama Masa Studi dengan melihat nilai tabulasi silang antar variabel. Berdasarkan Gambar 4, dapat kita lihat hasil diagram lingkaran dari variabel Lama Masa Studi.

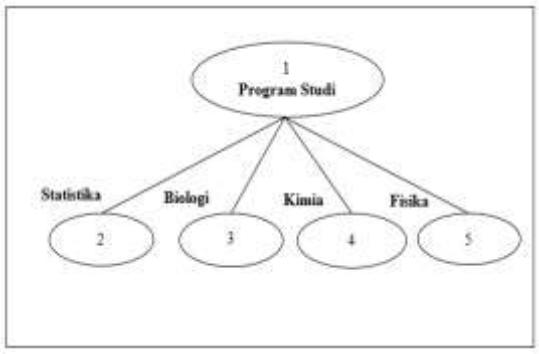
Algoritma C4.5

Pembentukan Pohon Klasifikasi

Tahap pertama dalam pembentukan pohon klasifikasi adalah pemilihan *node*akar. Perhitungan untuk menentukan *node*akar menggunakan Persamaan (1) untuk menentukan nilai *gain*dan Persamaan (2) untuk menentukan nilai *entropy*. Variabel yang digunakan untuk menentukan *node*akar adalah Indeks prestasi mahasiswa (IPK) (X_1), jenis kelamin (X_2), asal daerah (X_3), Program Studi (X_4), dan asal sekolah (X_5). Adapun hasil perhitungan *gain*berdasarkan Persamaan (2)dan *entropy* berdasarkan Persamaan (1)setiap variabel disajikan dalam Tabel 2

Tabel 2. Hasil Perhitungan *Gain*dan *Entropy* Untuk *Node*Akar

Node	Variabel	Kategori	Jumlah	> 5 tahun	≤ 5 tahun	*Entropy*	*Gain*
	Total		152	61	91	0,97171508	
1	IPK	IPK ≥ 3,50	43	7	36	0,640944185	0,0939351
		3,00 ≤ IPK < 3,50	99	46	53	0,996390619	
		IPK < 3,00	10	8	2	0,721928095	
	Jenis Kelamin	Perempuan	104	35	69	0,921467669	0,0270315
		Laki-laki	48	26	22	0,994984828	
	Asal Daerah	Samarinda	49	23	26	0,997294382	0,0065715
		Luar Samarinda	103	38	65	0,949848553	
	Program Studi	Statistika	37	16	21	0,98678672	0,1007336
		Biologi	39	19	20	0,999525689	
		Kimia	51	9	42	0,672294817	
		Fisika	25	17	8	0,904381458	
	Asal Sekolah	SMA/MA	140	54	86	0,96197806	0,0083246
		SMK	12	7	5	0,979868757	



Gambar 5. Hasil pembentukan cabang di variabel akar

Pada Gambar 5 dapat dilihat bahwa node 2, 3, 4, 5 membentuk node cabang karena data sampel masih berada pada masing-masing di dua kelas yaitu kelas > 5 tahun dan ≤ 5 tahun.

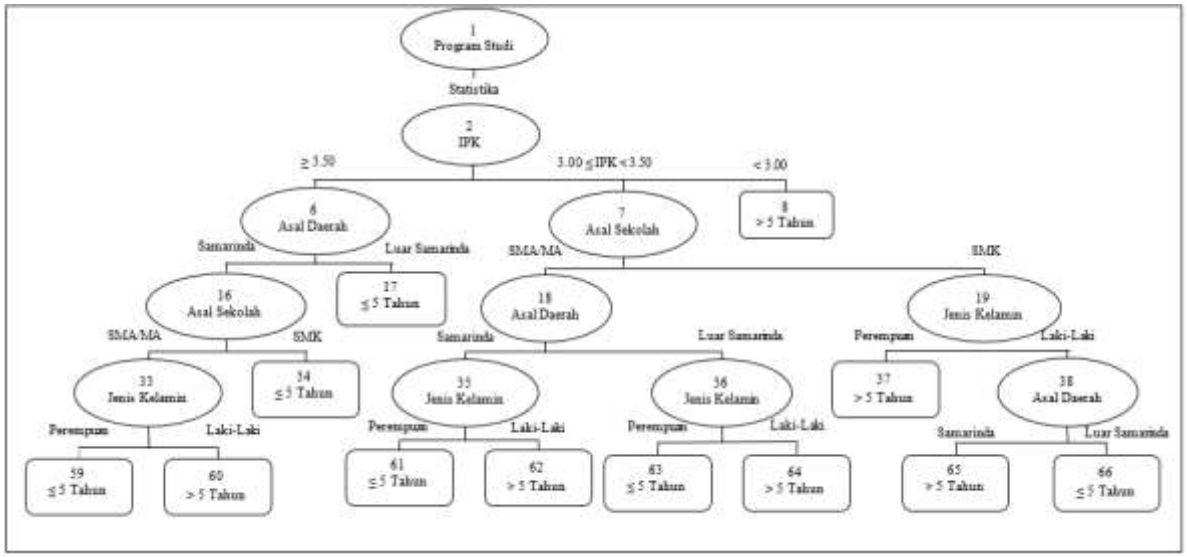
Pada Gambar 6 dapat dilihat bahwa pada program studi Statistika jika IPKnya masuk pada kategori 3,00 ≤ IPK < 3,50 berasal sekolah dari SMK, berjenis kelamin Laki-Laki dan berasal daerah dari Samarinda maka bisa diprediksi bahwa dia akan lulus > 5 tahun masih berada pada masing-masing di dua kelas yaitu kelas > 5 tahun dan ≤ 5 tahun.

Selanjutnya untuk *node* 2, *gain* dan entropy dihitung terlebih dahulu seperti pada langkah awal mencari *node* akar namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk dalam *node* 3, *node* 4 dan *node* 5. Perhitungan untuk menentukan untuk *node* selanjutnya menggunakan Persamaan (1) untuk menentukan nilai *gain* dan Persamaan (2) untuk menentukan entropy.Terus berlanjut sampai ke *node* 57 dimana semua *node* tidak adanya penurunan lagi. Karena tidak ada lagi *node* yang harus diproses, maka *decision tree* dinyatakan selesai. Hasil akhir *decision tree* disajikan pada Gambar 6.

Analisis *Classification and regression tree* (CART)

Dalam pembentukan pohon klasifikasi, terdapat 3 tahap yaitu pemilihan pemilah, penentuan terminal *node*, dan penandaan label kelas. Adapun data yang digunakan untuk proses pembentukan pohon klasifikasi ada 80% sampel (data learning) yaitu sebanyak 152 sampel, sedangkan 20% sampel sisanya untuk data testing pohon klasifikasi yang terbentuk yaitu sebanyak 38 sampel.

Pada penelitian ini metode pemilahan yang digunakan yaitu metode pemilahan Indeks *Gini* sesuai dengan persamaan (2.10). Klasifikasi pada penelitian ini dibagi menjadi dua kelas, yaitu C_1: Jika Lama Masa Studi ≤ 5 tahun = 91 mahasiswa, C_2: Jika Lama Masa Studi > 5 = 61 mahasiswa.
Variabel IPK mempunyai tiga kategori yaitu IPK ≥ 3,50, 3,00 ≤ IPK < 3,50 dan IPK < 3. Maka kemungkinan pemilah dari variabel ini adalah $2^{3-1} - 1 = 2^2 - 1 = 3$ pemilah yaitu {(IPK ≥ 3,50, 3,00 ≤ IPK < 3,50), (IPK < 3)}, {(3,00 ≤ IPK < 3,50, IPK < 3), (IPK ≥ 3,50)}, {(IPK ≥ 3,50, IPK < 3), (3,00 ≤ IPK < 3,50)}.



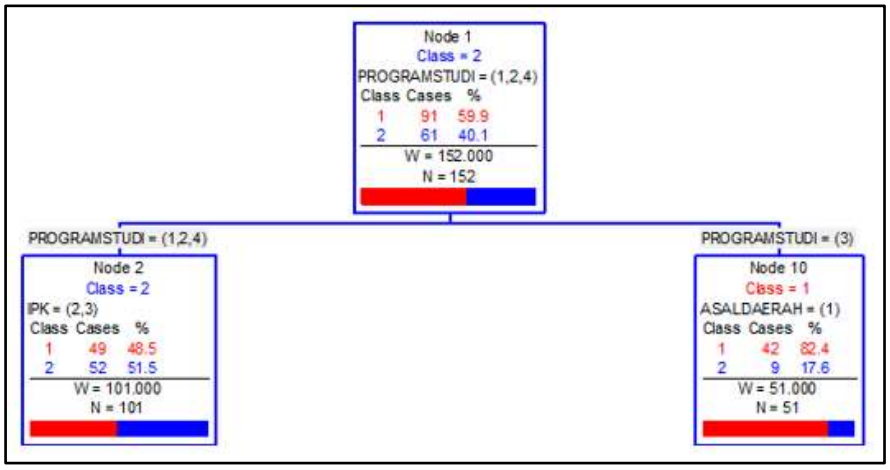
Gambar 6. Pohon Keputusan *node* 2 Statistika

Tabel 3. Pemilihan Pemilah Pada Setiap Variabel Prediktor

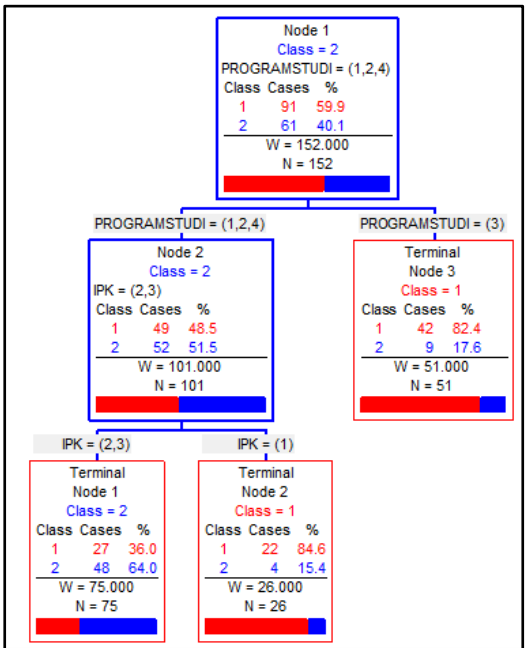
Pemilihian Pemilah	Kategori	*Gini*
IPK	{(IPK ≥ 3,50, 3,00 ≤ IPK < 3,50)} dan {(IPK < 3,00)}	0,4581
	{(3,00 ≤ IPK < 3,50, IPK < 3,00)} dan {(IPK ≥ 3,50)}	0,4356
	{IPK ≥ 3,50, IPK < 3,00)} dan {(3,00 ≤ IPK < 3,50)}	0,4655
Jenis Kelamin	Perempuan dan Laki-Laki	0,4623
Asal Daerah	Samarinda dan Luar Samarinda	0,4741
Program Studi	{Statistika, Biologi, Kimia} dan {Fisika}	0,4499
	{Statistika, Biologi, Fisika} dan {Kimia}	0,4295
	{Statistika, Kimia, Fisika} dan {Biologi}	0,4754
	{Biologi, Kimia, Fisika} dan {Statistika}	0,4799
	{Statistika, Biologi} dan {Kimia, Fisika}	0,4735
	{Statistika, Kimia} dan {Biologi, Fisika}	0,4427
	{Statistika, Fisika} dan {Biologi, Kimia}	0,4569
Asal Sekolah	SMA/MA dan SMK	0,4748

Dari hasil perhitungan nilai indeks *Gini* kelima variabel prediktor, dapat diketahui bahwa variabel yang memiliki nilai indeks *Gini* terkecil adalah variabel Program Studi pemilihan Program Studi kemungkinan ke dua yaitu dengan kategori {Statistika, Biologi, Fisika} dan {Kimia}, dengan nilai indeks *Gini*0,4295.Sehingga variabel Program Studi pemilihan Program Studi kemungkinan kedua dipilih sebagai pemilah pertama seperti Gambar 7.

Tahap selanjutnya adalah melakukan pohon klasifikasi maksimal. Proses pemangkasan pohon klasifikasi dimulai dengan mengambil t_R yang merupakan *node* anak kanan dan t_L yang merupakan *node* anak kiri dari T_{Max} yang dihasilkan oleh *node* induk *t*. Jika diperoleh dua *node* anak dan *node* induk yang memenuhi persamaan (2.14) maka *node* anak t_R dan t_L dapat dipangkas.Terdapat *node* yang akan dipangkas yaitu *node* pada kedalaman empat sampai dengan kedalaman 9. *Node* tersebut mengalami pemangkasan karena *node* induk dan *node*anaknya memenuhi persamaan (8).Setelah dilakukan proses pemangkasan sampai selesai, maka akan terbentuk suatu pohon optimal. Pohon klasifikasi optimal untuk data predikat kelulusan mahasiswa dapat dilihat pada gambar 8.



Gambar 7. Variabel Pemilah Pertama



Gambar 8. Pohon Klasifikasi Optimal Secara Rinci

Perbandingan Hasil Klasifikasi Algoritma C4.5 dan Algoritma CART

Berdasarkan analisis data yang telah dilakukan maka perbandingan ketepatan klasifikasi antara Algoritma C4.5 dan Algoritma CART diberikan pada Tabel4.

Tabel 4. Ketepatan Klasifikasi Kedua Metode

	C4.5	CART
APER	60 %	40 %
Ketepatan	40 %	60 %

Dari Tabel 4 dapat diketahui bahwa perbedaan hasil klasifikasi antara metode Algoritma C4.5 dan metode Algoritma CART terpaut jauh. Ketepatan klasifikasi baik dengan menggunakan metode Algortima C4.5 dan metode Algoritma CART, diperoleh nilai persentase ketepatan klasifikasi metode Algoritma C4.5 sebesar 40 % dan nilai persentase ketetapatan klasifikasi metode Algoritma CART sebesar 60 %. Berdasarkan nilai ketepatan klasifikasi, pada metode Algoritma C4.5 mempunyai ketepatan yang lebih rendah. Dengan demikian metode Algoritma CART merupakan metode yang lebih baik dibandingkan algoritma C4.5 dalam pengklasifikasian data Lama Masa Studi kelulusan seluruh mahasiswa FMIPA UNMUL tahun 2016.

Kesimpulan

Berdasarkan hasil analisis dan pembahasan yang dilakukan, diperoleh kesimpulan sebagai berikut :

1. Faktor-faktor yang berpengaruh pada Lama Masa Studi kelulusan seluruh mahasiswa FMIPA UNMUL tahun 2016 dengan menggunakan metode C4.5 adalah Program studi (X_4),Asal Sekolah (X_5) dan Asal Daerah (X_3).
2. Faktor-faktor yang berpengaruh pada Lama Masa Studi kelulusan seluruh mahasiswa FMIPA UNMUL tahun 2016 dengan menggunakan metode CART adalah Program studi (X_4) dan Indeks Prestasi Kumulatif (IPK) (X_1)
3. Hasil ketepatan klasifikasi pada metode algoritma C4.5 dan algoritma CART, diperoleh nilai persentase ketepatan klasifikasi metode algoritma C4.5 sebesar 40% dan nilai persentase ketepatan klasifikasi metode algoritma CART sebesar 60%. Dengan demikian metode algoritma CART merupakan metode yang lebih baik dibandingkan algoritma C4.5 dalam pengklasifikasian data Lama Masa Studi kelulusan seluruh mahasiswa FMIPA UNMUL tahun 2016.

Daftar Pustaka

Breiman, L., Friedman, J.H. Olsen, R.A. dan Stone, C.J. (1993). *Classification and regression tree*, New York: Chapman and Hall.

Efraim, Turban, J.E. Aronson, dan T.P. Liang. (2005). *Decision Support Systems and Intelligent System.* Yogyakarta: Andi Offset.

Fayyad, U. (1996). *Advances in Knowledge Discovery and Data Mining*. Yogyakarta: MIT Press.

Han J. and Kamber M. (2001). *Data Mining, Concepts and Techniques.* Simon Fraser University: Morgan Kaufmann Publisher.

Johnson, R.A.and Wichen, D.W. (2002). *Apllies Multivariate Statistical Analysis.* USA: Prentice-Hall.

Kusrini, dan E. T. Luthfy. (2009). *Algoritma Data Mining.* Yogyakarta: Andi Offset.

Larose, (2005). *Discovering Knowledge in Data*: *An Introduction to Data Mining.* New Jersey: John Willey and Sons.

Lewis, R.J. (2000). *An Introduction to Classification and regression tree Analysis*. Annual Meeting of the Society for Academic Emergency Medicine in San Fransisco. California: Departement of Emergency Medicine.

Mudyaharjo. (2002). *Filsafat ilmu pendidikan*. Bandung: Remaja Rosdakarya.

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

Santosa B. (2007). *Data Mining*: *Menambang Permata Pengetahuan di Gunung Data*. Yogyakarta: Graha Ilmu.

Statsoft (2003). *Classification and regression trees Theory and Application.*http://www.statsoft.com/textbook/stcart.html. Diakses pada tanggal 10 Februari 2017.

Tan, P. N., M. Steinbach., dan V. Kumar. (2006). *Introduction to Data Mining.* New York : Pearson Education.

Thomas, E. (2004). *Data Mining: Definition and Decision tree Examples,* USA : Pearson Education

Timofeev, R. (2004). *Classification an Regresion Trees Theory and Application.* A Master Thesis. Center Applied Statistics and Economics. Berlin: Humboldt University.

Yohannes, Y. dan Hoddinott, J. (1999). *Classification and regression tree: An Introduction.* Washington, D.C: International Food Policy Research Institute.