Aplikasi *Classification and Regression Tree* (CART) dan Regresi Logistik Ordinal dalam Bidang Pendididikan (Studi Kasus: Predikat Kelulusan Mahasiswa S1 Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman)

The Application of Classification and Regression Tree (CART) and Ordinal Logistic Regression in Education (Case Study: Predicate of Bachelor Degree's Graduation at Faculty of Mathematics and Natural Sciences)

David Siahaan[1], Sri Wahyuningsih[2], dan Fidia Deny Tisna Amijaya[3]
[1]Laboratorium Statistika Terapan FMIPA Universitas Mulawarman
[2,3] Program Studi Statistika FMIPA Universitas Mulawarman
Email: deviidsuho@gmail.com

Abstract

CART method is a nonparametric statistical methods which is for obtaining accurate data group in the classification analysis. CART main goal is to get an accurate data as a group identifier of a classification. CART can be applied in three main steps, namely the establishment of a classification tree, trimming the classification tree, and determination of optimal classification tree. Ordinal logistic regression is a statistical method for analysis response variables that have an ordinal scale consisting of three or more categories. Predictor variables that can be included in the model can be either continuous or categorical data consisting of two or more variables. This study wanted to know the classification results FMIPA UNMUL predicate graduation, the main factor that affect the predicate graduation FMIPA UNMUL who graduated in 2014, and a comparison of the accuracy of the classification results between CART and ordinal logistic regression. The results showed that gender (X_1), region origin (X_2), major (X_3), the status of secondary school (X_4), and duration of the study period (X_5) is the primary identifier graduation predicate FMIPA UNMUL, whereas gender (X_1) and duration of the study period (X_5) is a factor that affects the predicate graduation. Ordinal logistic regression model was able to predict with 65% accuracy, while the CART method has a predictive accuracy of 54.9%

Keywords: CART, classification trees, predicate graduation, ordinal logistic regression.

Pendahuluan

Pengklasifikasian merupakan salah satu metode statistik untuk mengelompokkan atau mengklasifikasikan suatu data yang disusun secara sistematis. Masalah klasifikasi sering dijumpai dalam kehidupan sehari-hari, baik pengklasifikasian data pada bidang akademik, kesehatan, segmentasi pasar, maupun pada bidang lainnya. Masalah-masalah tersebut dapat diselesaikan dengan metode klasifikasi, namun pada penyelesaiannya perlu diperhatikan dalam memilih metode klasifikasi yang tepat.

Metode klasifikasi dapat dilakukan dengan pendekatan parametrik dan nonparametrik. Salah satu metode klasifikasi dengan pendekatan nonparametrik yang sering digunakan adalah *Decision Tree* (Pohon keputusan). Pohon keputusan adalah suatu metode eksplorasi berstruktur pohon untuk melihat hubungan antar variabel respon dengan variabel penjelasnya. Beberapa metode yang dapat digunakan dalam metode pohon keputusan antara lain CHAID (*Chi-Squared Automatic Interaction Detection Analysis*), QUEST (*Quick, Unbiased Efficient, Statistical Tree*), CART(*Classification and Regression Tree*), dan lain-lain dimana masing-masing metode tersebut memiliki kekuatan dan kelemahannya masing-masing (Maimon and Rokach, 2010)

CART adalah salah satu metode atau alogaritma dari salah satu teknik eksplorasi data yaitu teknik pohon keputusan. CART dikembangkan untuk melakukan analisis klasifikasi pada variabel respon baik yang nominal, ordinal maupun kontinu. CART juga dapat menyeleksi variabel-variabel dan interaksi-interaksi variabel yang paling penting dalam menentukan hasil atau variabel prediktor (Breiman *et al*, 1993).

Regresi logistik terbagi menjadi tiga, yaitu analisis regresi logistik biner, regresi logistik nominal, dan regresi logistik ordinal. Regresi logistik biner digunakan ketika variabel prediktor terdapat dua kategori, regresi logistik nominal digunakan ketika variabelnya lebih dari dua. Sedangkan regresi logistik ordinal digunakan untuk menganalisis variabel respon yang mempunyai skala ordinal yang terdiri atas tiga kategori atau lebih.Variabel respon yang dapat disertakan dalam model berupa data kategori atau kontinu yang terdiri atas dua variabel atau lebih (Agresti, 2002).

Indeks Prestasi Kumulatif (IPK) adalah salah satu tolak ukur keberhasilan dalam studi seorang mahasiswa. Semakin tinggi IPK, maka mengindikasikan bahwa mahasiswa tersebut cerdas. Dan sudah menjadi kewajiban bagi universitas untuk mengontrol prestasi mahasiswanya. IPK akan digunakan untuk menentukan kriteria predikat kelulusan mahasiswa pada saat lulus nanti. Menurut buku peraturan akademik Universitas Mulawarman terdapat empat kategori predikat kelulusan mahasiswa, yaitu cukup, memuaskan, sangat memuaskan dan dengan pujian (*cumlaude*). Adapun tujuan penelitian ini yaitu untuk mengetahui hasil pengklasifikasian predikat kelulusan dengan menggunakan metode *CART*, mengetahui faktor yang mempengaruhi predikat kelulusan dengan regresi logistik ordinal, dan membandingkan ketepatan klasifikasi *CART* dan regresi logistik ordinal.

CART(*Classification and Regression Tree*)

CART adalah suatu metode atau algoritma dari salah satu teknik eksplorasi data yaitu teknik pohon keputusan. Metode ini di kembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olsen dan Charles J. Stone sekitar tahun 1980-an. CART dikembangkan untuk melakukan analisis klasifikasi pada variabel respon baik yang nominal, ordinal, maupun kontinu. CART juga dapat menyeleksi variabel-variabel dan interaksi-interaksi variabel yang paling penting dalam menentukan hasil atau variabel responnya. CART menghasilkan suatu pohon klasifikasi jika variabel responnya kategorik, dan menghasilkan pohon regresi jika variabel responnya kontinu. Tujuan utama CART adalah untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian (Timofeev, 2004). Pohon klasifikasi merupakan metode penyekatan data secara berulang (rekursif) dan secara biner (*binary recursive partitioning*), karena selalu membagi kumpulan data menjadi dua sekatan. Setiap sekatan data dinyatakan sebagai *node* dalam pohon yang terbentuk. Pembentukan pohon klasifikasi dilakukan melalui penyekatan gugus data dengan sederetan penyekat biner sampai dihasilkan *node* akhir. Proses pembentukannya terdiri dari 3 tahapan, yaitu pemilihan pemilah, penentuan *node* terminal, dan penandaan label kelas.

Pembentukan Pohon Klasifikasi

Tahap pertama membentuk pohon klasifikasi digunakan sampel data *Learning* (*L*) yang masih bersifat heterogen. Setiap pemilahan hanya bergantung pada nilai yang berasal dari suatu variabel independen. Rumus kemungkinan pemilah yaitu jika variabel prediktor kontinu = n – 1 pemilahan, jika variabel prediktor kategori nominal = 2^{L-1} – 1 pemilahan, dan jika variabel prediktor kategori ordinal = L – 1 pemilahan. Sampel tersebut akan dipilah berdasarkan aturan pemilahan dan kriteria *goodness-of-split*. Untuk mengukur tingkat keheterogenan suatu kelas dari suatu *node* tertentu dalam pohon klasifikasi dikenal dengan istilah *impurity measure i (t)*. Ukuran ini akan membantu kita menemukan fungsi pemilah yang optimal. Fungsi keheterogenan yang digunakan adalah indeks *Gini* seperti berikut:

$$i(t) = \sum_{j \neq i} p(j \mid t)\, p(i \mid t) \qquad (1)$$

$p(j \mid t)$ adalah peluang j pada *node t. Goodness of split* merupakan suatu evaluasi pemilahan oleh pemilah *s* pada *node t. Goodness of split* $¤(s,t)$ didefinisikan sebagai penurunan keheterogenan. Kualitas ukuran dari seberapa baik pemilah *s* dalam menyaring data menurut kelas merupakan ukuran penurunan keheterogenan dari suatu kelas dan didefinisikan sebagai

$$¤(s,t) = \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \qquad (2)$$

Tahap kedua adalah penentuan *terminal node*. Suatu *node t* akan menjadi *terminal node* atau tidak, akan dipilah kembali bila pada node *t* tidak terdapat penurunan keheterogenan secara berarti atau adanya batasan minimum *n* seperti hanya terdapat satu pengamataan pada tiap *node* anak. Umumnya jumlah kasus minimum dalam suatu terminal akhir adalah 5, dan apabila hal itu terpenuhi maka pengembangan pohon dihentikan (Lewis, 2000).

Tahap ketiga yaitu penandaan label kelas. Penandaan label kelas pada *terminal node* dilakukan berdasarkan aturan jumlah terbanyak, yaitu:

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

dimana $p(j \mid t)$ 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 *terminal node 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-tiap *node* anak atau adanya batasan minimum *n*, semua pengamatan dalam tiap *node* anak identik, dan adanya batasan jumlah level/kedalaman pohon maksimal.

Pemangkasan Pohon Klasifikasi

Setelah terbentuk pohon maksimal, tahap selanjutnya adalah pemangkasan pohon untuk mencegah terbentuknya pohon klasifikasi yang berukuran besar dan kompleks. Pemangkasan (*pruning*) yaitu suatu penilaian ukuran pohon tanpa mengorbankan ketepatan melalui

pengurangan *node* pohon sehingga mencapai ukuran pohon yang layak. Ukuran pemangkasan yang digunakan untuk memperoleh ukuran pohon yang layak adalah *Cost Complexity Minimum* (Breiman *et. al*, 1993). Ukuran *complexity* adalah sebagai berikut:

$$R_{\sim}(t) = R(T) + \sim \left|\tilde{T}\right| \quad (4)$$

dimana, $R(T)$ adalah *Resubtution Estimate* (Proporsi kesalahan pada sub pohon), α adalah kompleksitas Parameter (*Complexity Parameter*) dan $\left|\tilde{T}\right|$ adalah ukuran banyaknya *node* terminal pohon T.

Cost complexity prunning menentukan suatu pohon bagian $T(\sim)$ yang meminimumkan $R_{\sim}(t)$ pada seluruh pohon bagian, atau untuk setiap nilai $\sim$, dicari pohon bagian $T(\sim) < T\,max$ yang meminimumkan $R_{\sim}(t)$ yaitu:

$$R_{\sim}(T(\sim)) = min_{T<T\,max}\, R\sim(T) \quad (5)$$

Penentuan Pohon Klasifikasi Optimal

Setelah dilakukan pemangkasan diperoleh pohon klasifikasi optimal yang berukuran sederhana namun memberikan nilai pengganti yang cukup kecil. Penduga pengganti yang sering digunakan adalah validate silang lipat V (*Cross Validation V-Fold Estimates*)

Penduga validasi silang lipat V sering digunakan apabila amatan yang tidak cukup besar. Amatan dalam *L* dibagi secara aca menjadi bagian *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

$$R^{ts}(T_t^{(V)}) = \frac{1}{Nv}\sum_{(x_n j_n)\in L_V} X(d^{(V)}(x_n) \neq j_n) \quad (6)$$

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

$$R^{Cv}(T_t) = \frac{1}{V}\sum_{v=1}^{v} R^{ts}(T^{(v)}) \quad (7)$$

Regresi Logistik Ordinal

Regresi logistik ordinal merupakan salah satu metode statistik untuk mengganalisis variabel respon yang mempunyai skala ordinal yang terdiri atas tiga kategori atau lebih. Variabel prediktor yang dapat disertakan dalam model dapat berupa data kategori atau kontinu yang terdiri atas dua variabel atau lebih.

Model yang dapat dipakai untuk regresi logistik ordinal adalah model logit. Model logit tersebut adalah *cumulative logit models*. Pada model logit ini sifat ordinal dari respon Y dituangkan dalam peluang kumulatif yaitu peluang kurang dari satu atau sama dengan kategori respon ke-j pada *p* variabel prediktor yang dinyatakan dalam vektor X, $P(Y \le j|X)$, dengan peluang lebih besar dari kategori respon ke-j, $P(Y > j|X)$ (Hosmer dan Lameshow, 2000). Peluang kumulatif, $P(Y \le j|X)$, didefinisikan sebagai berikut:

$$P(Y \le j \mid X) = \frac{exp\left(\beta_j + \sum_{k=1}^{p} \beta_k X_k\right)}{1 + exp\left(\beta_j + \sum_{k=1}^{p} \beta_k X_k\right)} \quad (8)$$

dimana j = 1,2,…, J-1 adalah kategori respon (Agresti, 2002).

Estimasi Parameter

Estimasi parameter dapat dipergunakan metode maksimum *likelihood*. Metode ini memperoleh dugaan maksimum *likelihood* bagi β dengan langkah awal yaitu membentuk fungsi *likelihood*. Salah satu metode yang dapat digunakan dalam mengestimasi parameter model logistik adalah *Maximum Estimation Likelihood* (MLE). Pada dasarnya metode ini memberikan nilai estimasi parameter β dengan cara memaksimumkan fungsi *likelihood*-nya. Jika fungsi distribusi peluang untuk Y_i adalah $f(Y_i) = \gamma^{Y1}(1-\gamma)^{1-Y1}$, maka fungsi *likelihood* untuk n pengamatan bebas adalah:

$$L(\beta_0, \beta_1, \beta_2, \ldots, \beta_k) = \prod_{i=1}^{n}\left\{[¥(X_i)]^{Y_i}[1-¥(X_i)]^{1-Y_i}\right\}$$
$$= \left\{\left[\frac{¥(X_i)}{1-¥(X_i)^{i}}\right]^{\sum_{i=1}^{n} Y_i}[1-¥(X_i)]\right\} \quad (9)$$

Berdasarkan fungsi *likelihood* didapatkan ln fungsi *likelihood*nya sebagai berikut:

$$ln(L(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)) = \ell(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$$
$$= \sum_{i=1}^{n}\left\{Y_i(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k) + \left[ln1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}\right]\right\} \quad (10)$$

Estimasi parameter regresi logistik didapatkan dari turunan parsial pertama fungsi ln *likelihood* terhadap paramter yang akan diestimasi kemudian disamakan dengan nol.

Estimasi dari parameter regresi logistik ordinal didapatkan dengan menurunkan fungsi log *likelihood* terhadap parameter yang akan diestimasi dan disamakan dengan nol. Persamaan $\frac{\partial L(\quad)}{\partial \beta_k} = 0$ dipergunakan untuk estimasi parameter β_k dimana k=1, 2, …, n dan

$\frac{\partial L(\quad)}{\partial \quad_o} = 0$ merupakan estimasi intersep $\quad_o$ dimana j = 1, 2, …, j – 1.

Hasil persamaan $\frac{\partial L(\quad)}{\partial \quad_k} = 0$ dan $\frac{\partial L(\quad)}{\partial \quad_o} = 0$ merupakan fungsi nonlinier sehingga diperlukan metode iterasi untuk memperoleh estimasi parameternya. Metode iterasi yang dipergunakan adalah metode *iterative Weighted Least Square* (WLS) yaitu algoritma *Newton-Raphson.* (Agresti, 1990).

Uji Serentak

Uji serentak dilakukan dengan menggunakan uji G, yaitu pada dasarnya menunjukkan apakah semua variabel bebas yang dimasukkan dalam model mempunyai pengaruh secara bersama-sama terhadap variabel terikat. Adapun hipotesis yang digunakan adalah sebagai berikut:

Hipotesis

H_0 : $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (secara simultan variabel prediktor tidak berpengaruh terhadap variabel respon)

H_1 : $\beta i \neq 0$; i=1,2,…,k (minimal ada satu dari variabel prediktor yang berpengaruh terhadap variabel respon)

Taraf signifikansi

Taraf signifikansi yang digunakan adalah ~ = 0,05

Statistik uji

$$G = -2\frac{(n_1/n)^{n_0}}{\prod_{\quad_i}^{y_i}(1-\quad)^{1-y_i}} = -2\ln\frac{L_0}{L_k} \qquad (11)$$

Pengambilan Keputusan

Statistik uji G mengikuti distribusi *chi-Squared* dengan derajat bebas banyaknya parameter dalam model, karena itu untuk memperoleh keputusan uji adalah membandingkan nilai G dengan nilai

H_0 terima: jika $G \leq \chi^2_{(p,\alpha)}$ atau nilai *p-value* ≥ α

H_0 tolak : jika $G > \chi^2_{(p,\alpha)}$ atau nilai *p-value* < α

(Basuki, 2004)

Uji Individu

Untuk pengujian signifikansi parameter model secara individu dapat diuji dengan *Wald Test*. Hasil dari uji *Wald* ini akan menunjukkan apakah suatu variabel prediktor signifikan atau layak untuk masuk ke dalam model atau tidak.

Hipotesis

H_0 : $\beta_k = 0$, k = 1,2,…, n (tidak ada pengaruh variabel prediktor ke-k terhadap variabel respon)

H_1 : $\beta_k \neq 0$, k = 1,2,3,…,n (ada pengaruh variabel prediktor ke-k terhadap variabel respon)

Statistik Uji

$$W = \frac{\hat{\quad}_k}{SE(\hat{\quad}_k)} \qquad (12)$$

dimana,

$SE(\hat{\quad}_k) = \sqrt{var\ \hat{\quad}_k}$

W = Nilai statistik uji wald

$\hat{\quad}_k$ = Estimasi koefisien parameter ke –k

Daerah Kritis

H_0 ditolak bila $|W|$ lebih besar dari $Z_{\alpha/2}$ atau *p-value* kurang dari α. Hal ini dikarenakan statistik uji *W* mengikuti distribusi normal.

(Hosmer dan Lemeshow, 2000)

Uji Kecocokan Model Regresi Logistik Ordinal

Dalam mencocokan sebuah model logistik, perlu dipilih sebuah model dengan fungsi penghubung dan variabel penjelas yang hasilnya paling cocok. Uji ini digunakan untuk menilai kecocokan model dengan membandingkan hasil pengamatan dengan nilai dugaan.

Hipotesis

H_0 : Model sesuai (Tidak terdapat perbedaan antara hasil pengamatan dengan nilai dugaan)

H_1 : Model tidak sesuai (Terdapat perbedaan antar hasil pengamatan dengan nilai dugaan)

Statistik Uji

$$\hat{C} = \sum_{k=1}^{n}\frac{(o_k - n_k'\bar{\quad}_k)}{n_k'\bar{\quad}_k(1-\bar{\quad}_k)} \qquad (13)$$

dimana, $o_k = \sum_{k=1}^{C_a} y_i$

$$\bar{\quad}_k = \sum_{k=1}^{C_a}\frac{m_j\hat{\quad}_j}{n_k'}$$

n_k = total pengamatan grup k

Pengambilan keputusan

Uji ini mengikuti distribusi *Chi Squared* dengan derajat bebas *df – 2*. Daerah penolakan H_0 adalah jika nilai $\hat{C} > \chi^2_{(df-2)}$ atau nilai *P-value* < α.

(Hosmer and Lameshow, 2000)

Interpretasi Koefisien Model Regresi Logistik Ordinal

Interpretasi atau penaksiran dari perbandingan selisih/*odds ratio* (*ψ*) adalah menjelaskan berapa kali lipat kenaikan atau penurunan peluang Y = 1, jika nilai variabel prediktor (X) berubah sebesar nilai tertentu. Nilai *odds ratio* selalu positif. didapatkan penduga untuk *odds ratio* sebagai berikut:

$$\cdot = exp(\hat{\quad}_k) \qquad (14)$$

Indeks Prestasi Kumulatif

Penilaian keberhasilan akademik mahasiswa didasarkan pada nilai bobot rata-rata atau Indeks Prestasi (IP). Indeks Prestasi dibedakan atas Indeks Prestasi Semester dan Indeks Prestasi Kumulatif (IPK). IPK dan IPS dihitung dari mata kuliah yang tertera pada Kartu Hasil Studi (KHS).

Kriteria Predikat Kelulusan

Menurut buku peraturan akademik Universitas Mulawarman tahun 2014, IPK sebagai dasar penentuan predikat kelulusan Program Vokasi, Sarjana, dan Profesi adalah:

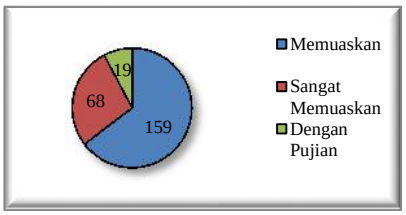
a. IPK 2,00 - 2,75 : Cukup
b. IPK 2,76 – 3,50 : Memuaskan
c. IPK 3,51 – 3,69 : Sangat memuaskan
d. IPK ≥ 3,70 : Dengan pujian (*cum laude*), jika mahasiswa dapat menyelesaikan masa studi tidak melebihi n + 0,5. Tidak pernah mengulang mata kuliah dan tanpa nilai C serta semua mata kuliah ditempuh di UNMUL.

Metodologi Penelitian

1. Analisis Deskriptif
   Pada analisis ini menggunakan bantuan software *SPSS 17* dengan menyajikan grafik dari data predikat kelulusan mahasiswa FMIPA UNMUL tahun 2014.
2. Analisis CART
   Dalam analisis menggunakan metode CART ada beberapa langkah-langkah yaitu sebagai berikut:
   a. Membentuk pohon klasifikasi, yang terdiri dari pemilahan pemilah terbaik *n* penentuan *terminal node*. Pemilahan terbaik dihitung berdasarkan selisih terbesar rata-rata kuadrat sisa antara antara *node* induk dan kedua *node* anak yang memilahnya ketika tidak memungkinkan lagi melakukan pemilahan pada suatu *node* atau jika tidak terdapat penurunan keheterogenan yang berarti, sehingga tidak akan dipilah lagi.
   b. Memangkas pohon klasifikasi yang paling kecil dengan menggunakan kriteria kompleksitas kesalahan (*cost complexity*) yang minimum
   c. Memilih pohon terbaik
   d. Hasil klasifikasi dan interpretasi
3. Analisis regresi logistik ordinal
   Dalam analisis menggunakan metode regresi logistik ordinal ada beberapa langkah yang harus dilakukan yaitu:
   a. Pengujian serentak. Melakukan uji untuk mengetahui ada tidaknya pengaruh antara variabel respon dan variabel prediktor secara.
   b. Pengujian individu. Melakukan pengujian secara individu untuk mengetahui ada tidaknya pengaruh antara variabel respon dan variabel prediktor.
   c. Pemilihan model regresi logistik ordinal terbaik.
   d. Interpretasi model regresi logistik ordinal terbaik.

Hasil dan Pembahasan

a. Statistika Dekriptif



Gambar 1. *Pie Chart* untuk Predikat Kelulusan

Berdasarkan gambar dapat diketahui bahwa mahasiswa yang lulus dengan predikat memuaskan ada 159 mahasiswa. Mahasiswa yang lulus dengan predikat sangat memuaskan ada 68 mahasiswa dan mahasiswa yang lulus dengan predikat dengan pujian ada 19 mahasiswa dari total sebanyak 246 mahasiswa lulusan program Sarjana FMIPA UNMUL.

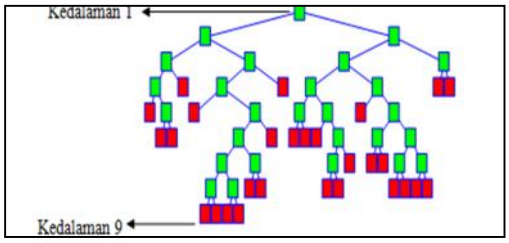
b. Analisis CART

Tahap pertama pembentukan pohon klasifikasi maksimal adalah pemilihan pemilah. Perhitungan pemilah setiap variabel prediktor diperoleh hasil seperti berikut

a. Variabel jenis kelamin mempunyai dua kategori yaitu perempuan dan laki-laki. Maka banyak kemungkinan pemilah dari variabel ini adalah $2^{2-1}-1 = 1$ pemilah.
b. Variabel asal daerah memiliki dua kategori, yaitu Samarinda dan luar Samarinda. Maka banyak kemungkinan pemilah variabel ini adalah $2^{2-1}-1 = 1$ pemilah.
c. Variabel program studi (prodi), mempunyai 5 kategori yaitu Biologi, Fisika, Ilkom, Kimia, Statistika. Maka banyak kemungkinan pemilah dari variabel ini adalah $2^{5-1}-1 = 15$ pemilah.
d. Variabel status sekolah menengah memiliki dua kategori, yaitu SMA dan SMK. Maka banyak kemungkinan pemilah variabel ini adalah $2^{2-1}-1 = 1$ pemilah.
e. Variabel lama masa studi memiliki dua kategori, yaitu lulus ≤ 5 tahun dan lulus antara 5 sampai 7 tahun. Maka banyak kemungkinan pemilah variabel ini adalah $2^{2-1}-1 = 1$ pemilah

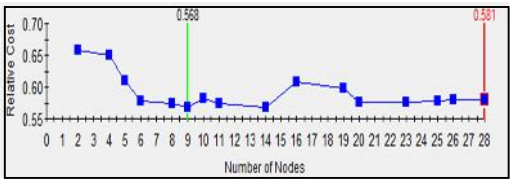
Perhitungan pemilah menggunakan rumus indeks Gini. Berdasarkan nilai Indeks *Gini* kelima variabel prediktor, dapat diketahui bahwa variabel yang memiliki nilai indeks *Gini* terkecil adalah variabel lama masa studi dengan nilai indeks *Gini* 0,4578. Sehingga variabel lama masa studi dipilih

sebagai pemilah pertama. Setelah semua pemilah ditentukan maka kita dapatkan pohon klasifikasi maksimal pada Gambar 2.



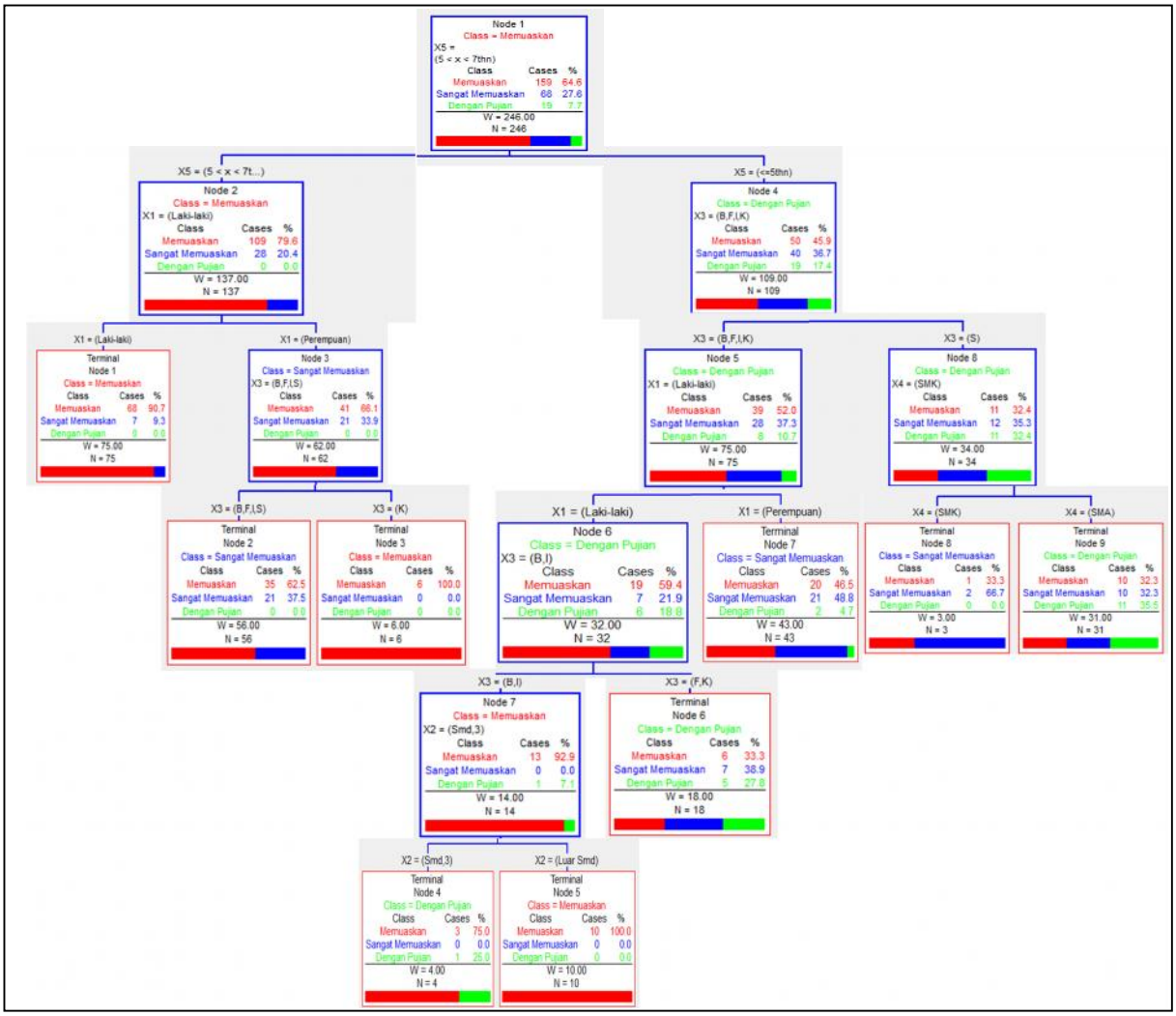
Gambar 2. Pohon Klasifikasi Maksimal

Tahap selanjutnya adalah melakukan pemangkasan 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 dari *node* induk *t*. Proses ini diulangi sampai tidak ada lagi pemangkasan yang mungkin terjadi, sehingga diperoleh ukuran pohon yang layak dan memenuhi *cost complexity minimum* seperti pada Gambar 3.



Gambar 3. Plot *Relative Cost*

Garis hijau pada Gambar 4 menunjukkan nilai *relative cost minimum* pada pohon optimal sebesar 0,568 dan garis merah menunjukkan nilai *relative cost maximum* pada pohon maksimal sebesar 0,581. Sedangkan nilai *test set relative cost* dan *parameter complexity* masing-masing sebesar 0,56799 ± 0,05904. Gambar 4 memberikan informasi bahwa nilai *relative cost* pohon klasifikasi maksimal lebih besar dibandingkan dengan nilai *relative cost* pohon klasifikasi optimal. Oleh karena itu harus dilakukan pemangkasan pohon maksimal agar didapatkan nilai *relative cost* yang paling kecil. Setelah proses pemangkasan selesai maka akan terbentuk pohon klasifikasi optimal pada Gambar 4.



Gambar 4 Pohon Klasifikasi Optimal

Berdasarkan Gambar 4 dapat diketahui bahwa kelima variabel prediktor masuk menjadi pemilah pembentuk pohon klasifikasi optimal. Artinya kelima variabel prediktor tadi merupakan kelompok penciri data dari klasifikasi predikat kelulusan FMIPA UNMUL tahun 2014.

Ketepatan Klasifikasi

Pohon klasifikasi optimal yang telah terpilih tadi kemudian diuji tingkat keakuratannya dalam mengelompokkan data, dapat kita lihat ketepatan klasifikasi sebagai berikut:

Tabel 1. Tabel Ketepatan Klasifikasi

Aktual	Kelas Prediksi		
	Memuaskan	Sangat Memuaskan	Dengan Pujian
Memuaskan	82	49	28
Sangat Memuaskan	14	38	16
Dengan Pujian	1	3	15

Dari Tabel 1 diperoleh ketepatan pengklasifikasian sebesar:

$$\frac{14+1+49+3+28+16}{246}=\frac{111}{246}=0{,}451$$

Sehingga ketepatan klasifikasinya adalah 1 – 0,451 = 0,549 atau 54,9%. Artinya pohon klasifikasi yang terbentuk mampu memprediksi dengan tepat pengamatan sebesar 54,9%.

c. Regresi Logistik Ordinal

Adapun langkah-langkah pengujian regresi logistik ordinal yaitu, uji serentak, uji individu, dan uji kecocokan model regresi logistik ordinal.

Uji Serentak

Uji ini berfungsi untuk mengetahui apakah variabel prediktor mempunyai pengaruh yang signifikan terhadap variabel respon secara keseluruhan.

Hipotesis

H_0 : $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (secara serentak tidak ada pengaruh yang signifikan antara variabel prediktor terhadap variabel respon)

H_1 : Minimal ada satu $\beta_i \neq 0$, i = 1, 2, …k (secara serentak ada pengaruh yang signifikan antara variabel prediktor terhdapa variabel respon)

Statistik Uji

Tabel 2. Uji Simultan atau Uji *Likelihood Ratio*

G	DF	P_value	Log-Likelihood
49,176	8	0,000	-180,627

Kriteria Pengujian

H_0 diterima jika $G \leq \chi^2_{(p,\alpha)}$ atau nilai P-*value* $\geq \alpha$, H_0 ditolak jika $G > \chi^2_{(p,\alpha)}$ atau nilai P-*value* $< \alpha$

Keputusan

Dari tabel diperoleh nilai G (49,176) > $\chi^2_{(8,0,05)}$ (15,51) dan nilai P-*value* (0,000) < 0,05. Maka H_0 ditolak.

Kesimpulan

Dari hasil analisis dengan menggunakan uji G *likelihood ratio* dapat disimpulkan bahwa minimal ada satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon.

Uji Individu

Uji ini berfungsi untuk mengetahui apakah ada pengaruh dari variabel prediktor terhadap variabel respon secara individu. Uji ini menggunakan statistik uji dari *Wald* sehingga bisa juga disebut sebagai uji *Wald*.

Tabel 3. Uji Individu

Variabel	*W*	P-Value	Keputusan
Jenis Kelamin (X_1)	2,49	0,013	H_0 ditolak
Asal Daerah (X_2)	0,63	0,526	H_0 diterima
Program Studi (X_3)	1,32	0,188	H_0 diterima
Status Sekolah Menengah (X_4)	0,94	0,947	H_0 diterima
Lama Studi (X_5)	4,35	0,000	H_0 ditolak

Dari pengujian secara individu dapat kita lihat bahwa variabel prediktor yang berpengaruh signifikan terhadap predikat keluluan mahasiswa FMIPA UNMUL adalah variabel jenis kelamin dan variabel lama studi.

Uji Kecocokan Model Regresi Logistik Ordinal

Uji ini berfungsi untuk menilai kesesuaian model regresi logistik dengan membandingkan hasil pengamatan dengan nilai dugaan.

Hipotesis

H_0 : Model sesuai (tidak terdapat perbedaan antara hasil pengamatan dengan nilai dugaan)

H_1 : Model tidak sesuai (terdapat perbedaan antara hasil pengamatan dengan nilai dugaan)

Statistik Uji

Tabel 4. Uji kecocokan model

Method	Chi-Squared	Df	P_value
Pearson	6,681	4	0,154

Keputusan

Jika $\hat{C} < ¡^2(df-2)$ atau nilai P-value > 0,05 maka H_0 diterima

Jika $C \geq ¡^2(df-2)$ atau nilai P-value ≤ 0,05 maka H_0 ditolak

Kesimpulan

Dari Tabel 4.25 diketahui bahwa nilai $\hat{C} = 6{,}681 < ¡^2_{(4)} = 9{,}49$ dan nilai *P value* adalah sebesar 0,154 > α = 0,05 maka H_0 diterima. Sehingga dapat disimpulkan bahwa model sesuai atau tidak terdapat perbedaan antara hasil pengamatan dengan nilai dugaan.

Model Regresi Logistik Ordinal

Berdasarkan hasil pengujian simultan maupun parsial didapatkan model regresi logistik ordinal dengan 2 intersep karena pada variabel respon terdapat 3 kategori (Y=1, 2 dan 3)

Tabel 5. Hasil Estimasi Parameter

Variabel	Keterangan	W	P-value	Coef(β)
	Const (1)	-2,33	0,020	-0,501
	Const (2)	6,01	0,000	1,601
X_1	Jenis Kelamin	2,81	0,005	0,834
X_2	Lama studi	5,18	0,000	1,502

Model regresi logistik ordinal

$$P(Y \leq 1 \mid X) = \sim_1 + {}_1X_1 + {}_2X_2$$
$$= -0{,}501 + 0{,}834\,X_1 + 1{,}502\,X_5$$
$$P(Y \leq 2 \mid X) = \sim_2 + {}_1X_1 + {}_2X_2$$
$$= 1{,}601 + 0{,}834\,X_1 + 1{,}502\,X_5$$

Atau model peluang persamaan logistiknya adalah

$$P(Y \leq 1 \mid X) = \frac{exp(-0{,}501 + 0{,}834X_1 + 1{,}502X_5)}{1 + exp(-0{,}501 + 0{,}834X_1 + 1{,}502X_5)}$$

$$P(Y \leq 2 \mid X) = \frac{exp(1{,}601 + 0{,}834X_1 + 1{,}502X_5)}{1 + exp(1{,}601 + 0{,}834X_1 + 1{,}502X_5)}$$

Ketepatan Klasifikasi Regresi Logistik Ordinal

Berdasarkan model regresi logistik ordinal dapat kita tentukan hasil klasifikasi. Untuk melihat keakuratan model regresi logistik ordinal dalam mengklasifikasikan data maka digunakanlah tabel ketepatan klasifikasi

Tabel 5. Tabel Ketepatan Klasifikasi Model Regresi Logistik Ordinal

Aktual	Prediksi		
	Memuaskan	Sangat Memuaskan	Dengan Pujian
Memuaskan	128	31	0
Sangat Memuaskan	36	32	0
Dengan Pujian	8	11	0

Dari tabel 5 diperoleh ketepatan pengklasifikasian sebesar:

$$\frac{36+8+31+11}{246} = \frac{86}{246} = 0{,}35$$

Sehingga ketepatan klasifikasinya adalah 1 – 0,35 = 0,65 atau 65%. Artinya model regresi logistik ordinal mampu memprediksi dengan tepat pengamatan sebesar 65%.

Interpretasi Koefisien Model Regresi Logistik Ordinal

Berdasarkan model regresi logistik ordinal dapat kita tentukan nilai *Odds Ratio* seperti di bawah ini:

- Untuk variabel jenis kelamin

  $\cdot = exp(\hat{}_k) = exp(0{,}834) = 2{,}30$

  Jadi, mahasiswa dengan jenis kelamin perempuan memiliki peluang 2,30 kali lebih besar untuk mendapatkan predikat kelulusan dengan pujian daripada mahasiswa berjenis kelamin laki-laki.

- Untuk variabel lama masa studi

  $\cdot = exp(\hat{}_k) = exp(1{,}502) = 4{,}49$

  Jadi, mahasiswa dengan lama masa studi ≤ 5 tahun memiliki peluang 4,49 kali lebih besar untuk mendapatkan predikat kelulusan dengan pujian daripada mahasiswa dengan lama masa studi antara 5 sampai 7 tahun.

Kesimpulan

Berdasarkan hasil analisis dan pembahasan yang dilakukan, kesimpulan yang diperoleh dari penelitian mengenai predikat kelulusan mahasiswa program sarjana FMIPA UNMUL yaitu:

1. Berdasarkan hasil klasifikasi CART, variabel prediktor yang menjadi penciri utama variabel predikat kelulusan mahasiswa program sarjana FMIPA UNMUL adalah variabel jenis kelamin, asal daerah, program studi, status sekolah menengah dan lama masa studi.
2. Berdasarkan analisis regresi logistik ordinal, faktor-faktor yang mempengaruhi predikat kelulusan mahasiswa program sarjana FMIPA UNMUL adalah variabel jenis kelamin dan lama masa studi.
3. Berdasarkan tabel ketepatan klasifikasi, model regresi logistik ordinal lebih baik dalam memprediksi hasil pengamatan daripada metode CART. Hal ini dilihat dari tingkat keakuratan klasifikasi model regresi logistik ordinal yang bernilai 65%, sedangkan metode CART hanya memiliki keakuratan klasifikasi sebesar 54,9%.

Daftar Pustaka

Agresti, A. 2002. *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.

Basuki, Achmad. 2004. Modeling dan Simulasi. Surabaya: IPTAQ Mulia Media

Breiman, L., Friedman, J.H., Olsen, R.A., dan Stone, C.J. 1993. *Classification and Regression Trees*. New York: Chapman & Hall.

Hosmer, D. W., and Lameshow, S. 2000. *Applied Logistic Regression*. New York: John Wiley & Sons, Inc.

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

Maimon, Oded and Rokach, Lior. 2010. *Data Mining and Knowledge Discovery Handbook*. Springer.

Timofeev, Roman. 2004. *Classification and Regression Trees (C&RT) Theory and Application*. A Master Thesis. CASE-Center of Applied Statistics and Economics. Berlin: Humboldt University.