Penerapan Metode Analisis Regresi Logistik Biner Dan *Classification And Regression Tree* (CART) Pada Faktor yang Mempengaruhi Lama Masa Studi Mahasiswa

Application Of Binary Logistic Regression And Classification And Regression Tree (CART) Methods On Factors Affecting Length Of Study Period Of Students

Chairunnisa[1], Yuki Novia Nasution[2], dan Ika Purnamasari[3]
[1]Laboratorium StatistikaTerapan FMIPA Universitas Mulawarman
[2,3]Jurusan Matematika FMIPA Universitas Mulawarman
[1]E-mail: annisa.nadhifa24@gmail.com

Abstract

Binary Logistic Regression is one of the logistic regression analysis which is used to analyze the relationship between a dichotomous dependent variable with several independent variables. Classification and Regression Tree (CART) is one of the methods that developed to perform classification analysis on dependent variables either on nominal, ordinal, or continuous scale. In this research, Binary Logistic Regression method and Classification and Regression Tree (CART) applied to the data of the students at Faculty of Math and Natural Science Mulawarman University graduated in year 2016 to determine the characteristic of student which is classified according to two categories that is the study period less than or equal to 5 years and study period more than 5 Years, with five independent variables namely GPA Graduates (X_1)*, Gender* (X_2)*, Type of Junior School* (X_3)**,** *Domicile* (X_4)*, and Major* (X_5)*. Factors that influence the study period of the students based on Binary Logistic Regression method are GPA, Gender, Secondary School Type and Major. The result of classification by using CART method is the student who have the study period less than or equal to 5 Years is a student from Chemistry major or have GPA between 3,51 and 4,00, while the study period more than 5 Year is the student who have GPA between 2,00 and 2,75; 2,76 and 3,50. In terms of classification accuracy, Binary Logistic Regression method was able to accurately predict the observation as much as 75.0%, while the CART method was able to accurately predict the observation as much as 77.27%.*

Keywords **:** *CART, Accuracy of Classification, Study Period, Binary Logistic Regression.*

Pendahuluan

Pendidikan penting bagi setiap orang sebagai bekal untuk dapat melangsungkan kehidupannya. Pentingnya pendidikan bagi setiap orang di dalam sebuah negara akan memberikan pengaruh positif terhadap negara tersebut karena dengan pendidikan akan meningkatkan kualitas sumber daya manusia sehingga bagi negara tentu akan menambah daya saing terhadap negara lain. Berdasarkan Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 49 Tahun 2014 tentang Standar Nasional Perguruan Tinggi yang tertulis pada Pasal 17 Nomor 3 masa studi terpakai bagi mahasiswa dengan beban belajar program Diploma IV atau Sarjana adalah selama 5 tahun.

Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA) Universitas Mulawarman merupakan salah satu fakultas dengan jumlah mahasiswa yang relatif banyak dan memiliki empat Jurusan antara lain Statistika, Kimia, Biologi, dan Fisika**.** Pada kenyataannya, cukup banyak mahasiswa yang masa studinya panjang dikarenakan banyak faktor-faktor yang mempengaruhi ketidaktepatan waktu kelulusan mahasiswa tersebut. Sehingga perlu dilakukan analisis untuk mengetahui faktor apa saja yang mempengaruhi lama masa studi mahasiswa FMIPA.

Regresi logistik biner merupakan salah satu regresi logistik yang digunakan untuk menganalisis hubungan antara satu variabel terikat dan beberapa variabel bebas, dengan variabel terikatnya berupa data kualitatif dikotomi yaitu bernilai 1 untuk menyatakan keberadaan sebuah karakteristik dan bernilai 0 untuk menyatakan ketidakberadaan sebuah karakteristik (Hosmer & Lemeshow, 2000).

CART adalah salah satu metode atau algoritma dari salah satu teknik eksplorasi data yaitu teknik pohon keputusan. CART dikembangkan untuk melakukan analisis klasifikasi pada variabel terikat baik yang nominal, ordinal maupun kontinu (Breiman dkk, 1993).

Berdasarkan hal-hal yang telah diuraikan, maka Penulis tertarik untuk membahas tentang penggunaan Metode Regresi Logistik Binerdan *Classification and Regression Tree* (CART) dalam menganalisis lama masa studi mahasiswa Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA) Universitas Mulawarman dengan variabel-variabel yang mempengaruhi lama masa studi mahasiswa.

Regresi Logistik Biner

Analisis Regresi Logistik Biner memiliki tujuan untuk mendapatkan model terbaik dan sederhana yang menggambarkan hubungan antara variabel terikat dengan variabel-variabel bebas, dengan variabel terikatnya yang bersifat dikotomi dan variabel bebasnya dapat berupa variabel kualitatif dan kuantitatif(Hosmer dan Lemeshow, 2000).

Syarat utama dalam regresi logistik biner adalah variabel terikatnya berupa variabel biner yaitu variabel diskrit dengan dua nilai. Misalnya diambil ilustrasi, bila variabel terikat tidak terjadi diberi nilai 0 dan bila variabel terikat terjadi diberi nilai 1, sedangkan variabel bebasnya dapat berupa variabel kuantitatif. Untuk nilai variabel bebas bertipe kualitatif biasanya disebut juga variabel boneka (*dummy*) dan dilakukan pemberian suatu angka agar dapat dianalisis. Variabel kualitatif ini berupa variabel dikotomi atau dapat berupa variabel polikotomus. Untuk variabel berupa kuantitatif didefinisikan secara langsung dan biasanya disebut variabel kontinu.

Penaksiran Parameter Regresi Logistik Biner

Dalam Hosmer dan Lemeshow (2000), penaksiran parameter pada model regresi logistik yang mempunyai variabel terikat dikotomi adalah menggunakan metode *Maximum Likelihood Estimation* (MLE). Pada dasarnya metode MLE menetapkan asumsi distribusi Bernoulli dan objek pengamatan saling bebas atau memberikan nilai taksiran parameter dengan memaksimumkan fungsi *likelihood* (*likelihood function*). Dalam bentuk persamaan matematis, persamaan logistik dinyatakan dalam bentuk berikut :

$$\pi(x)=\frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}} \qquad (1)$$

Fungsi ini merupakan probabilitas dari data dalam menghasilkan nilai estimasi parameter.

Jika y=1, maka $P(y=1\mid x)=\frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}}$ dan ,

Jika y=0, maka $P(y=0\mid x)=\frac{1}{1+e^{\left[\beta_0+\sum_{j=1}^{p}\beta_j x_j\right]}}$

Bentuk fungsi *likelihood*nya adalah :

$$L(\beta)=\prod_{i=1}^{n}\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right]^{y_i}\left[1-\pi(x_i)\right] \qquad (2)$$

dengan $L(\beta)$ = *likelihood.*

Untuk memaksimalkan fungsi *likelihood*, rumus tersebut diubah ke dalam bentuk log *likelihood* dengan notasi $l(\beta)$ untuk memudahkan penyelesaian persamaan matematisnya dapat diperoleh persamaan :

$$l(\beta)=\sum_{i=1}^{n}\{y_i\ln[\pi(x_i)]+(1-y_i)\ln[1-\pi(x_i)]\} \qquad (3)$$

Pengujian Parameter

Pengujian parameter dalam regresi penting untuk dilakukan. Hal ini dikarenakan pengujian tersebut digunakan untuk menentukan apakah variabel bebas dalam model signifikan terhadap variabel terikat. Pengujian dapat dilakukan secara:

Uji Simultan

Menurut Hosmer dan Lemeshow (2000), uji simultan bertujuan untuk mengetahui pengaruh variabel bebas secara serentak atau simultan terhadap variabel terikat. Langkah pengujiannya adalah sebagai berikut :

Hipotesis :

H_0 : $\beta_1=\beta_2=\cdots=\beta_P=0$

H_1 : Paling sedikit ada satu $\beta_j\neq 0$, (minimal ada satu variabel bebas yang berpengaruh secara simultan terhadap variabel terikat) dimana $j=1,2,\ldots,p$

Statistik Uji :

Statistik uji yang digunakan adalah G, yakni *likelihood ratio* :

$$G=2\left\{\sum_{i=1}^{n}[y_i\ln(\hat{\pi}_i)+(1-y_i)\ln(1-\hat{\pi}_i)]-[n_1\ln(n_1)+n_0\ln(n_0)-n\ln(n)]\right. \qquad (4)$$

Statistik uji G ini mengikuti distribusi *chi-square* dengan db adalah derajat bebas yang merupakan banyaknya variabel bebas dalam model. Keputusan uji diperoleh dengan membandingkan G dan nilai χ^2 (H_0 ditolak jika nilai G $\geq$ $\chi^2_{(\alpha,db)}$).

Uji Parsial

Dalam Hosmer dan Lemeshow (2000), uji parsial dilakukan dengan menguji setiap β_j secara individual akan menunjukan apakah suatu variabel bebas layak untuk masuk dalam model atau tidak.

Hipotesis :

H_0:$\beta_j=0$, (tidak ada pengaruh variabel bebas ke-j terhadap variabel terikat)

H_1: $\beta_j\neq 0$, (ada pengaruh variabel bebas ke-j terhadap variabel terikat) dimana j = 1, 2, ... ,p

Statistik Uji :

$$W=\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)},\text{dengan } SE(\hat{\beta}_j)=\sqrt{\text{var}(\hat{\beta}_j)} \qquad (5)$$

Statistik uji Wald mengikuti distribusi normal sehingga memperoleh keputusan pengujian, nilai W dibandingkan dengan nilai $Z_{\alpha/2}$ (H_0 ditolak jika nilai W $\geq Z_{\alpha/2}$ atau *p-value*$\leq \alpha$).

Uji Kesesuaian Model

Dalam Hosmer dan Lemeshow (2000), uji statistik *Goodness of Fit* digunakan untuk mengetahui kesesuaian model. Langkah Uji *Goodness of Fit* adalah sebagai berikut :

Hipotesis :

H_0: Tidak ada perbedaan antara hasil pengamatan dengan nilai dugaan

H_1: Ada perbedaan antara hasil pengamatan dengan nilai dugaan

Statistik Uji :

$$\hat{C} = \sum_{k=1}^{g} \frac{\left(O_k - n_k' \bar{\pi}_k\right)^2}{n_k' \hat{\pi}_k \left(1 - \bar{\pi}_k\right)} \qquad (6)$$

Jika H_0 benar, maka distribusi statistik uji $\hat{C}$ mengikuti distribusi *chi-square* dengan derajat bebas g-2.Daerah penolakan H_0 adalah jika $\hat{C} \geq \chi^2_{(\alpha,g-2)}$ atau *p-value*$\leq \alpha$.

Penafsiran Koefisien Model Regresi Logistik

Dalam menginterpretasikan atau menafsirkan koefisien β_j pada regresi logistik, hal yang harus selalu diperhatikan adalah jenis variabel bebasnya, berupa dikotomi, polikotomus atau kontinu. *Odds ratio* (ψ) digunakan untuk mengintrepretasi model regresi logistik.

Untuk variabel bebas yang bersifat dikotomi, diasumsikan nilai x adalah 0 dan 1, sehingga dalam model akan terdapat dua nilai $\pi(x)$ dan dua nilai $1 - \pi(x)$.

Tabel 1. Nilai Model Regresi Logistik untuk Variabel Bebas bersifat Biner

		Variabel Bebas (x)	
		$x = 1$	$x = 0$
Variabel terikat (y)	$y = 1$	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
	$y = 0$	$1-\pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0) = \frac{1}{1+e^{\beta_0}}$

(Sumber : Hosmer dan Lemeshow, 2000)

Odds ratio untuk terjadinya variabel terikat di antara variabel bebas yang mempunyai nilai x=1 adalah :

$$\frac{P(y=1\,|x=1)}{P(y=0\,|x=1)} = \frac{\pi(1)}{1-\pi(1)}$$

Sedangkan *Odds ratio* untuk terjadinya variabel terikat di antara variabel bebas yang mempunyai x=0 adalah :

$$\frac{P(y=1\,|x=0)}{P(y=0\,|x=0)} = \frac{\pi(0)}{1-\pi(0)}$$

Odds ratio adalah rata-rata besarnya kecenderungan variabel terikat bernilai tertentu jika x=1 dibandingkan dengan x=0, dilambangkan dengan ψ dan dinyatakan dalam persamaan :

$$\psi = \frac{\frac{\pi(1)}{(1-\pi(1))}}{\frac{\pi(0)}{(1-\pi(0))}} = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))},$$

Interpretasi *odds-ratio* (ψ) adalah menjelaskan beberapa kali lipat kenaikan atau penurunan peluang y=1, jika nilai variabel bebas (X) berubah sebesar nilai tertentu nilai *odds ratio* selalu positif.

Log *odds ratio* merupakan perbedaan atau selisih nilai logistik.

Dengan mensubsitusikan model regresi logistik pada Tabel 1, maka *odds ratio* menjadi :

$$\psi = e^{\beta_1}$$

Sehingga log *odds ratio* menjadi : $\ln(\psi) = \beta_1$

(Hosmer dan Lemeshow, 2000)

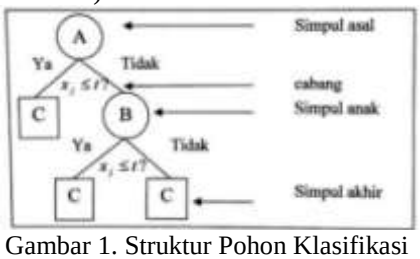
Untuk variabel bebas yang bersifat polikotomus, metode yang dipilih adalah menggunakan koleksi desain variabel (*Dummy variabel).*

***Classification and Regression Tree* (CART)**

CART adalah salah satu metode atau algoritma dari salah satu teknik eksplorasi data yaitu teknik pohon keputusan. Metode ini dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen dan Charles J. Stone sekitar Tahun 1980-an. CART menghasilkan suatu pohon klasifikasi jika variabel terikatnya kategorik, dan menghasilkan pohon regresi jika variabel terikatnya kontinu. Tujuan utama CART adalah untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian (Timofeev, 2004).

Menurut Breiman, dkk (1993), keunggulan dari CART adalah tidak perlu dipenuhinya asumsi sebaran oleh semua variabel, serta algoritmanya yang langsung dapat menangani masalah data hilang. CART juga tidak dipengaruhi oleh pencilan, kolinieritas, heteroskedastisitas yang biasanya mempengaruhi metode parametrik. Dalam CART, pencilan akan diisolasi ke dalam *node* (simpul) tertentu sehingga tidak mempengaruhi penyekatan, karena menggunakan variabel yang kolinier sebagai penyekat buatan.

Pohon klasifikasi merupakan metode penyekatan data secara berulang (rekursif) dan secara biner, karena selalu membagi kumpulan data menjadi dua sekatan. Setiap sekatan data dinyatakan sebagai simpul (*node*) dalam pohon yang terbentuk. Hasil dari proses penyekatan ini direpresentasikan dalam suatu struktur pohon seperti terlihat pada Gambar 1 (Breiman dkk,1993). Lewis (2000) menyebut simpul asal (*root*) sebagai simpul induk (*parent node*), simpul induk dapat disekat menjadi simpul anak (*children node*).



Gambar 1. Struktur Pohon Klasifikasi

Pada Gambar 1, A, B dan C merupakan variabel-variabel penjelas yang terpilih untuk menjadi simpul. A merupakan simpul asal atau simpul induk, sementara B dan C merupakan simpul anak dimana C juga merupakan simpul akhir yang tidak bercabang lagi. Sementara *t* merupakan suatu nilai yang merupakan nilai tengah antara dua nilai amatan variabel x_j secara berurutan. Adapun beberapa langkah-langkah penerapan dalam CART, yaitu Pembentukan

Pohon Klasifikasi, Pemangkasan Pohon Klasifikasi, dan Penentuan Pohon Klasifikasi

Pembentukan Pohon Klasifikasi

Proses pembentukanpohonklasifikasi terdiri atas 3 tahapan, yaitu pemilihan pemilah, penentuan simpul terminal, dan penandaan label kelas.

Pemilihan Pemilah (Classifier)

Untuk membentuk pohon klasifikasi digunakan sampel data *Learning* (L) yang masih bersifat heterogen. Rumus kemungkinan pemilah disajikan sebagai berikut :

Variabel bebas kontinu = $n - 1$ pemilahan (7.a)

Variabel bebas kategori nominal= $2^{L-1} - 1$ (7.b)

Variabel bebas kategori ordinal = $L - 1$ (7.c)

dimana :

n = Banyaknya data pada satu variabel bebas

L = Data *Learning*

Sampel tersebut akan dipilah berdasarkan aturan pemilahan dan kriteria *goodness of split*. Pemilihan pemilah bergantung pada jenis pohon atau lebih tepatnya tergantung pada jenis variabel terikatnya. Untuk mengukur tingkat keheterogenan suatu kelas dari suatu simpul tertentu dalam pohon klasifikasi dikenal dengan istilah *impurity measure i(t)*. Metode fungsi *impurity measurei (t)* yang sering digunakan adalah Indeks *Gini* :

$$i(t) = \sum_{j \neq i} p(j|t)p(k|t) \qquad (8)$$

dimana :

k, j = Kelas

$p(j|t)$ = Probabilitas bersyarat kelas j yang berada pada simpul t

Goodness Of Split merupakan suatu evaluasi pemilihan oleh pemilah *s* pada simpul *t. Goodness Of Split* $\phi(s,t)$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 :

$$\phi(s,t) = \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \qquad (9)$$

dengan :

$i(t)$ = Fungsi keheterogenan indeks *Gini*

$\phi(s,t)$ = Kriteria *Goodness Of Split*

$p_L i(t_L)$ = Proporsi pengamatan dari simpul t menuju simpul kiri

$p_R i(t_R)$ = Proporsi pengamatan dari simpul t menuju simpul kanan

Pemilah yang menghasilkan nilai $\Delta i(s,t)$ lebih tinggi merupakan pemilah yang lebih baik karena hal ini memungkinkan untuk mereduksi keheterogenan secara lebih signifikan. Partisi t_L dan t_R, merupakan partisi dari simpul t menjadi dua himpunan bagian saling lepas dimana $p_L$dan p_R adalah proporsi masing-masing peluang simpul. Karena $t_L \cup t_R = t$ maka nilai $\Delta i(s,t)$ mempresentasikan perubahan dari keheterogenan dalam simpul *t* yang semata-mata disebabkan oleh pemilah *s*. Pengembangan pohon dilakukan dengan mencari semua kemungkinan pemilah pada simpul t_1 sehingga ditemukan pemilah s^* yang memberikan nilai penurunan keheterogenan tertinggi yaitu :

$$\Delta i(s^*, t_1) = max_{s \in S} \Delta i(s, t_1) \qquad (10)$$

(Breiman dkk, 1993).

Penentuan Simpul Terminal

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

Penandaan Label Kelas

Penandaan label kelas pada simpul terminal dilakukan berdasarkan aturan jumlah terbanyak, yaitu :

$$p(j_o|t) = max_j \; p(j|t) = max_j \; \frac{N_j(t)}{N(t)} \qquad (11)$$

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

Proses pembentukan pohon klasifikasi berhenti saat terdapat hanya satu pengamatan dalam tiap-tiap simpul anak atau adanya batasan minimum *n* , semua pengamatan dalam tiap simpul anak identik, dan adanya batasan jumlah level/kedalaman pohon maksimal. Setelah terbentuk pohon maksimal, tahap selanjutnya adalah pemangkasan pohon untuk mencegah terbentuknya pohon klasifikasi yang berukuran besar dan kompleks.

Indeks *Gini* (*Gini Index)*

Aturan pemisahan *Gini* (indeks *Gini*) adalah aturan yang paling luas atau paling sering digunakan. Aturan ini mengikuti fungsi *impurity i(t),* dimana fungsi *impurity* adalah suatu fungsi yang digunakan untuk mengukur keakuratan model dengan memberikan indikasi kehomogenan kelas-kelas pada data sehingga pada simpul akhir (*terminal node*) akan didapatkan data yang lebih murni. Berdasarkan persamaan (8), rumus Indeks *Gini* adalah :

$$i(t) = \sum_{j \neq k} p(j|t)p(k|t)$$

Indeks *Gini* memiliki interpretasi yang menarik. Sebagai ganti penggunaan aturan pluralitas untuk mengklasifikasikan objek dalam *t* simpul, menggunakan aturan yang memberikan suatu objek yang dipilih secara acak dari simpul

(*node*) ke kelas j dengan probabilitas $p(j|t)$. Probabilitas memperkirakan bahwa item sebenarnya di kelas *k* adalah $p(k|t)$.

Jika simpul *t* dibelah menjadi 2 buah subset *L* dan *R* dengan ukuran masing-masing $n_1$dan n_2, indeks *gini* dari pembelahan tersebut didefinisikan sebagai berikut :

$$gini_{pembela\ han}(t) = \frac{n_1}{n}\, i(t_L) + \frac{n_2}{n}\, i(t_R) \quad (12)$$

dimana :

$gini_{pembela\ han}$ = Nilai indeks *Gini* setiap variabel
$i(t_L)$ = Nilai indeks *Gini* subset *L*
$i(t_R)$ = Nilai indeks *Gini* subset *R*
n = Banyaknya seluruh data
n_1 = Banyaknya data pada subset *L*
n_2 = Banyaknya data pada subset *R*

Indeks *Gini* ini sederhana dan cepat dihitung, indeks ini juga dapat menggabungkan nilai kesalahan variabel simetri dengan cara alami.

Pemangkasan Pohon Klasifikasi

Untuk mendapatkan pohon yang layak maka perlu dilakukan pemangkasan (*pruning*) yaitu suatu penelitian ukuran pohon tanpa mengorbankan ketepatan melalui pengurangan simpul pohon sehingga dicapai ukuran pohon yang layak. Dengan kata lain pemangkasan pohon dilakukan untuk mendapatkan pohon akhir yang lebih sederhana. Ukuran pemangkasan yang digunakan untuk memperoleh ukuran pohon yang layak adalah *Cost Complexity Minimum* (Breiman dkk, 1993).

Sebagai ilustrasi, untuk sembarang pohon *T* yang merupakan subpohon dari pohon terbesar T_{max} diperoleh ukuran *complexity*-nya yaitu $|\tilde{T}|$ yang menyatakan banyaknya simpul akhir pada pohon *T* tersebut. Dalam regresi biasa, $|\tilde{T}|$ analog dengan derajat bebas model. Untuk suatu $\alpha \geq 0$, yang disebut sebagai parameter *complexity* yakni ukuran tentang *cost* bagi penambahan satu simpul akhir pada pohon *T*, maka besarnya *Resubtitution* estimasi pohon T pada kompleksitas α adalah :

$$R_\alpha(t) = R(T) + \alpha|\tilde{T}| \quad (13)$$

dimana :

$R_\alpha(t)$ = *Resubtitution* suatu pohon T pada kompleksitas α
$R(T)$ = *Resubtitution Estimate* (Proporsi kesalahan pada sub pohon)
α = Kompleksitas Parameter (*Complexity Parameter*)
$|\tilde{T}|$ = Ukuran banyaknya simpul terminal pohon T(*complexity)*

Resubtitution suatu pohon T pada kompleksitas αmerupakan kombinasi linier dari *Resubtitution Estimate* dan nilai kompleksitas pohon yang dibentuk. *Cost complexity prunning* menentukan suatu pohon bagian $T(\alpha)$ yang meminimumkan $R_\alpha(t)$ pada seluruh pohon bagian untuk setiap nilai α. Nilai kompleksitas α akan secara perlahan meningkat selama proses pemangkasan. Selanjutnya mencari pohon bagian max $(T(\alpha) < T)$ yang meminimumkan $R_\alpha(t)$ yaitu :

$$R_\alpha(T(\alpha)) = \min_{T<T_{MAX}} R_\alpha(T) \quad (14)$$

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).* Yang perlu diperhatikandalam pemangkasan bukan T_{max} melainkan T_1, yakni suatu sub pohon yang memenuhi kriteria $R(T_1) = R(T_{max})$.

Untuk mendapatkan T_1dari T_{max}, ambil t_L dan t_R yang merupakan simpul anak kiri dan simpul anak kanan dari T_{max} yang dihasilkan dari simpul induk *t.* Kemudian hitung nilai t_L dan t_R menggunakan persamaan

$$R(T) = r(t)P(t) \quad (15)$$

Dengan nilai $r(t) = 1 - max_j\ P(j|t)$ dan $P(j|t)$ adalah peluang beberapa objek yang berada dalam simpul, begitu diperoleh dua simpul anak dan simpul induknya yang memenuhi persamaan

$$R(T) = R(t_L) + R(t_R) \quad (16)$$

maka pangkas simpul anak t_L dan t_R tersebut (Breiman dkk, 1993).

Penentuan Pohon Klasifikasi Optimal

Pohon 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 independent (*independent test sample estimate*) dan penduga validasi silang lipat V (*cross validation V-fold estimate*).

Penduga validasi silang lipat V sering digunakan apabila pengamatan yang dilakukan tidak cukup besar atau sampel kurang dari 3000. Pengamatan 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 :

$$R^{ts}(T_t^{(V)}) = \frac{1}{Nv}\sum_{(x_n j_n)\in L_v} X(d^{(V)}(x_n) \neq j_n) \quad (17)$$

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

$$R^{Cv}(T_t) = \frac{1}{V}\sum_{v=1}^{V} R^{ts}(T^{(v)}) \quad (18)$$

Pohon klasifikasi optimum dipilih T^* dengan

$$R^{Cv}(T^*) = \min_t R^{Cv}(T_t) \quad (19)$$

Ketepatan Klasifikasi

Kriteria perbandingan teknik klasifikasi didasarkan pada kesalahan klasifikasi yang dikenal dengan *Apparent Error Rate* (APER) merupakan nilai dari besar kecilnya jumlah observasi yang salah dalam pengklasifikasian berdasarkan suatu fungsi klasifikasi (Johnson dan Wichen, 2007).

Tabel 2. Klasifikasi Aktual dan Prediksi

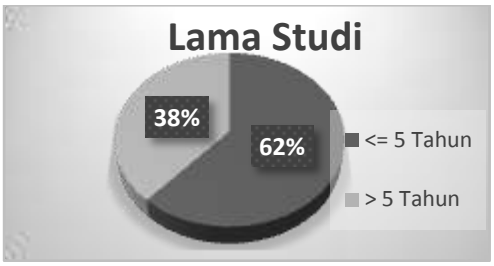
Kelas Aktual	Prediksi Kelas	
	y=0	y=1
y=0	n_{00}	$n_{01=}$ n_0- n_{00}
y=1	$n_{10=}$ n_1- n_{11}	n_{11}

$$APER = \frac{n_{00}+n_{11}}{n_0+n_1} \times 100\% (20)$$

Hasil dan Pembahasan

1. Analisis Statistika Deskriptif

Karakteristik yang digambarkan pada analisis deskriptif adalah IPK, jenis kelamin, jenis sekolah menengah, daerah asal, jurusan



Gambar 2. Persentase Lama Studi

Gambar 2 menunjukkan lama studi mahasiswa FMIPA Universitas Mulawarman Tahun 2016. Dari 188 mahasiswa, sebesar 62 % atau 116 mahasiswa FMIPA Universitas Mulawarman masuk dalam kategori 1 yaitu lama studi $\leq$ 5 Tahun.

2. Analisis Regresi Logistik Biner

Model regresi logistik biner sesuai pada persamaan (1) dengan 5 variabel bebas diperoleh hasil estimasi parameter model regresi sebagai berikut :

$$\pi(x) = \frac{e^{-1{,}199-1{,}809\,X_{1(1)}-21{,}688\,X_{1(2)}-1{,}207\,X_2+1{,}828\,X_3-0{,}144\,X_4+X_{5(1)}+2{,}368\,X_{5(2)}+0{,}126\,X_{5(3)}}}{1+e^{-1{,}199-1{,}809\,X_{1(1)}-21{,}688\,X_{1(2)}-1{,}207\,X_2+1{,}828\,X_3-0{,}144\,X_4+X_{5(1)}+2{,}368\,X_{5(2)}+0{,}126\,X_{5(3)}}}$$

Tabel 3 menunjukkan Pengkodean Desain Variabel Polikotomus. Berdasarkan Tabel 3 yang menjadi IPK Mahasiswa respondennya adalah $3{,}51 \leq$ IPK $\leq 4{,}00$ kedua desain variabel yang terbentuk yaitu D_1 dan D_2 keduanya akan sama dengan 0, ketika IPK dari responden adalah $2{,}76 \leq$ IPK $\leq 3{,}50$ D_1 akan sama dengan 1 dan D_2 masih sama dengan 0; ketika IPK dari respondennya adalah $2{,}00 \leq$ IPK $\leq 2{,}75$, maka D_1 akan sama dengan 0 dan D_2 = 1.

Tabel 3.Variabel Rancangan Untuk IPK Mahasiswa

IPK Mahasiswa	Variabel Rancangan	
	D_1	D_2
$\mathbf{3,51 \leq IPK \leq 4,00}$	0	0
$\mathbf{2,76 \leq IPK \leq 3,50}$	1	0
$\mathbf{2,00 \leq IPK \leq 2,75}$	0	1

Tabel 4.Variabel Rancangan Untuk Jurusan

Jurusan	Variabel Rancangan		
	D_1	D_2	D_3
Biologi	0	0	0
Fisika	1	0	0
Kimia	0	1	0
Matematika	0	0	1

Berdasarkan Tabel 4 yang menjadi Jurusan respondennya adalah Biologi ketiga desain variabel yang terbentuk yaitu D_1 , D_2 dan D_3 ketiganya akan sama dengan 0 ; ketika Jurusan dari responden adalah FisikaD_1 akan sama dengan 1, D_2 masih sama dengan 0 dan D_3 masih sama dengan 0 ; ketika Jurusan dari respondennya adalah Kimia maka D_1 akan sama dengan 0 dan D_2 sama dengan 1 dan D_3 samadengan 0 ; ketika Jurusan dari respondennya adalah Matematika maka D_1 akan sama dengan 0 dan D_2 sama dengan 1 dan D_3 sama dengan 1.

Pengujian Signifikansi Parameter

Uji Simultan

Hipotesis

H_0 : $\beta_{1(1)} = \beta_{1(2)} = \beta_2 = \beta_3 = \beta_4 = \beta_{5(1)} = \beta_{5(2)} = \beta_{5(3)} = 0$ (secara simultan variabel bebas tidak berpengaruh terhadap variabel terikat).

H_1 :Paling sedikit ada satu $\beta_j \neq 0$, dengan j = 1,2,...,5 (minimal ada satu variabel bebas yang berpengaruh secara simultan terhadap variabel terikat).

Taraf Signifikansi

$\alpha = 0{,}05$ atau 5 %

Statistik Uji

Uji G (*Likelihood Ratio Test*) pada Persamaan (4), dimana :

$n_1 = 116$ (Mahasiswa Lama Studi $\leq$ 5 Tahun)

$n_0 = 72$ (Mahasiswa Lama Studi $>$ 5 Tahun)

$n = 188$ (Jumlah Mahasiswa Lulus Tahun 2016)

Perhitungan :

Berdasarkan perhitungan menggunakan *software SPSS 20* diperoleh :

G = 358,962

Daerah Penolakan

Menolak H_0 jika nilai G $\geq \chi^2_{(0,05;5)} = 11{,}07$

Keputusan dan Kesimpulan

Berdasarkan hasil perhitungan diperoleh nilai G = 358,962 dimana nilai G = 358,962 $\geq \chi^2_{(0,05;5)} = 11,07$ maka dapat putuskan bahwa menolak H_0 sehingga dapat disimpulkan bahwa minimal ada satu variabel bebas yang berpengaruh terhadap variabel terikat.

Uji Parsial

Tabel 5. Uji Parsial Variabel Konstanta dan Variabel Bebas

Var	$\hat{\beta}_j$	Wald	*p-value*	Keputusan
$\mathbf{X_{1(1)}}$	-1,809	10,775	0,001	Menolak H_0
$\mathbf{X_{1(2)}}$	-21,68	0,000	1,000	Menerima H_0
$\mathbf{X_2}$	1,207	7,483	0,006	Menolak H_0
$\mathbf{X_3}$	1,828	4,746	0,029	Menolak H_0
$\mathbf{X_4}$	-1,144	0,135	0,713	Menerima H_0
$\mathbf{X_{5(1)}}$	0,000	0,000	1,000	Menolak H_0
$\mathbf{X_{5(2)}}$	2,368	19,601	0,000	Menolak H_0
$\mathbf{X_{5(3)}}$	0,126	0,058	0,809	Menerima H_0
C	-1,199	1,223	0,269	Menerima H_0

Hipotesis

H_0 : $\beta_0 = 0$ (tidak ada pengaruh variabel konstanta terhadap variabel terikat).

H_1 : $\beta_0 \neq 0$ (ada pengaruh variabel konstanta terhadap variabel terikat)

Taraf Signifikansi

$\alpha = 0,05$ atau 5 %

Statistik Uji

Uji Wald pada Persamaan (5)

Daerah Penolakan

Menolak H_0 jika nilai p-value $\leq \alpha$ atau nilai$W \geq Z_{\alpha/2}$

Keputusan dan Kesimpulan

Berdasarkan Tabel 5 diperoleh nilai *p-value* untuk variabel konstanta adalah sebesar 0,269, dan nilai Uji Wald = 1,223<$Z_{0,05/2}$ = 1,960 maka dapat disimpulkan bahwa tidak ada pengaruh variabel konstanta terhadap variabel terikat.

Hipotesis

H_0 : $\beta_j = 0$ (tidak ada pengaruh variabel bebas ke-j terhadap variabel terikat).

H_1 : $\beta_j \neq 0$ (ada pengaruh variabel bebas ke-j terhadap variabel terikat) dimana j = 1,2,...,5.

Taraf Signifikansi

$\alpha = 0,05$ atau 5 %

Statistik Uji

Uji Wald pada Persamaan (5)

Daerah Penolakan

Menolak H_0 jika nilai p-value $\leq \alpha$ atau nilai $W \geq Z_{\alpha/2}$

Keputusan dan Kesimpulan

Berdasarkan Tabel 6 diperoleh nilai *p-value* untuk variabel $X_{1(1)}$, $X_{1(2)}$, X_2, X_3, X_4, $X_{5(1)}$, $X_{5(2)}$, $X_{5(3)}$, masing-masing adalah 0,001 ; 1,000 ; 0,006 ; 0,029 ; 0,713 ; 1,000 ; 0,000 ; dan 0,809 dan nilai Uji Wald adalah 10,775 ; 000 ; 7,483 ; 4,746 ; 0,135 ; 0,000 ; 19,601 ; 0,058 maka dapat disimpulkan bahwa variabel yang berpengaruh secara parsial terhadap lama studi adalah IPK ($X_{1(1)}$), Jenis Kelamin (X_2), Jenis Sekolah Menengah (X_3) dan Jurusan ($X_{5(2)}$).

Model Terbaik Regresi Logistik Biner

Berdasarkan hasil output *SPSS 20* model terbaik dengan menggunakan metode *Backward Wald* diperoleh tabel sebagai berikut :

Tabel 6. Nilai Model Regresi Logistik

Variabel	$\hat{\beta}_j$	*P-value*
$\boldsymbol{X_{1(1)}}$	-1,820	0,001
$\boldsymbol{X_2}$	-1,221	0,006
$\boldsymbol{X_3}$	1,816	0,030
$\boldsymbol{X_{5(2)}}$	2,359	0,000

Diperoleh model sebagai berikut :

$$\pi(x) = \frac{e^{\beta_{1(1)}X_{1(1)}+\beta_2X_2+\beta_3X_3+\beta_{5(2)}X_{5(2)}}}{1+e^{\beta_{1(1)}X_{1(1)}+\beta_2X_2+\beta_3X_3+\beta_{5(2)}X_{5(2)}}}$$

Tabel 6 menunjukkan bahwa model sudah baik. Karena semua variabel memiliki nilai *p-value* < 0,05. Sehingga dapat disimpulkan bahwa faktor-faktor yang sangat berpengaruh pada lama studi adalah IPK ($X_{1(1)}$) , Jenis Kelamin (X_2) , Jenis Sekolah Menengah (X_3) dan Jurusan ($X_{5(2)}$). Sehingga model regresi logistik terbaik untuk memprediksi kepuasan pelanggan adalah:

$$\pi(x) = \frac{e^{-1,820\,X_{1(1)}-1,221\,X_2+1,816\,X_3+2,359X_{5(2)}}}{1+e^{-1,820\,X_{1(1)}-1,221\,X_2+1,816\,X_3+2,359X_{5(2)}}}$$

Uji Kesesuaian model

Hipotesis

H_0 : Model sudah sesuai (Tidak ada perbedaan antara hasil pengamatan dengan hasil dugaan)

H_1 : Model belum sesuai (Ada perbedaan antara hasil pengamatan dengan hasil dugaan)

Taraf Signifikansi

$\alpha = 0,05$ atau 5 %

Statistik Uji

Uji *Goodness of Fit* pada Persamaan (6)

Daerah Penolakan

Menolak H_0 jika *p-value* ≤0,05 atau nilai $\hat{C} \geq \chi^2_{(\alpha,g-2)}$

Keputusan dan Kesimpulan

Berdasarkan Tabel Uji Hosmer dan Lemeshow, menunjukkan bahwa nilai p-value = 0,478 >$\alpha = 0,05$dan nilai $\hat{C} = 6,545 < \chi^2_{(0,05,7)} = 14,067$ maka dapat diputuskan bahwa gagal menolak H_0 sehingga dapat disimpulkan bahwa tidak ada perbedaan antara hasil pengamatan dengan nilai dugaan atau model regresi tersebut layak untuk digunakan.

Interpretasi Model

Berikut merupakan tabel nilai *odds ratio* masing-masing variabel.

Tabel 7 Kontribusi Variabel *X* Terhadap *Y*

Variabel	Exp($\hat{\beta}_j$)
$X_{1(1)}$	0,162
X_2	3,390
X_3	6,145
$X_{5(2)}$	10,583

Tabel 7 menjelaskan bahwa :

1. Mahasiswa yang nilai IPK 2,76 ≤ IPK ≤ 3,50 memiliki peluang sebesar 0,162 kali lebih kecil menempuh Lama Studi ≤ 5 Tahun dibandingkan dengan mahasiswa yang nilai IPK adalah 3,51 ≤ IPK ≤ 4,00 .
2. Mahasiswa dengan jenis kelamin Perempuan memiliki peluang sebesar 3,390 kali lebih lebihbesar menempuh Lama Studi ≤ 5 Tahun dibandingkan dengan dengan mahasiswa dengan jenis kelamin Laki-Laki.
3. Mahasiswa dengan jenis sekolah menengah SMA memiliki peluang sebesar 6,145 kali lebih besar untuk menempuh Lama Studi ≤ 5 Tahun dibandingkan dengan dengan mahasiswa dengan jenis sekolah menengah SMK.
4. Mahasiswa yang berasal dari jurusan Kimia memiliki peluang sebesar 10,583 kali lebih besar menempuh Lama Studi ≤ 5 Tahun dibandingkan dengan mahasiswa yang berasal dari jurusan Biologi.

Ketepatan Klasifikasi Model

Berikut adalah hasil perhitungan ketepatan klasifikasi model :

$$\frac{53+88}{188} \times 100\% = 75\ \%$$

Berdasarkan perhitungan ketepatan klasifikasi diperoleh 75 % yang artinya pohon klasifikasi yang terbentuk mampu memprediksi dengan tepat pengamatan sebesar 75% .

3. ***Classification and Regression Tree*** **(CART)**

Pembentukan Pohon Klasifikasi

Dalam proses pembentukan pohon klasifikasi, terdapat 3 tahapan yaitu pemilihan pemilah, penentuan simpul terminal, dan penandaan label kelas.

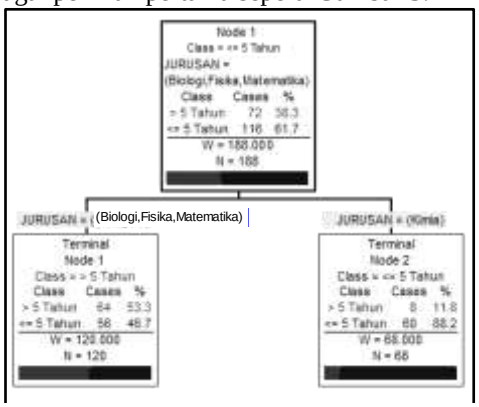
Pemilihan Pemilah

Perhitungan pemilah pada setiap variabel yaitu dengan menggunakan persamaan (7.b) untuk variabel penjelas kategori nominal yaitu variabel Jenis Kelamin (X_2) , Jenis Sekolah Menengah (X_3), Daerah Asal (X_4) , dan Jurusan (X_5), dan persamaan (7.c) untuk variabel penjelas kategori ordinal yaitu Variabel IPK (X_1). Adapun pemilihan pemilah dilakukan dengan menghitung nilai *IndeksGini*setiapkemungkinanpemilah.

Tabel 8. Rekapitulasi Index *Gini* untuk 5 Variabel Bebas

Variabel	Nilai Index *Gini*
IPK kemungkinan ke-1	0,4260
IPK kemungkinan ke-2	0.4685
Jenis Kelamin	0,4528
Jenis Sekolah Menengah	0,4504
Asal Daerah	0,4723
Jurusan kemungkinan ke-1	0,4572
Jurusan kemungkinan ke-2	0,4418
Jurusan kemungkinan ke-3	**0,3928**
Jurusan kemungkinan ke-4	0,4714
Jurusan kemungkinan ke-5	0,4538
Jurusan kemungkinan ke-6	0,4444
Jurusan kemungkinan ke-7	0,4122

Dari hasil perhitungan nilai indeks *Gini* kelima variabel bebas, dapat diketahui bahwa variabel yang memiliki nilai indeks *Gini* terkecil adalah variabel Jurusan dengan nilai indeks *Gini* 0,3928. Sehingga variabel Jurusan kemungkinan ke-3 yaitu Biologi, Fisika, dan Matematika dipilih sebagai pemilah pertama seperti Gambar 3.



Gambar 3. Variabel Pemilah Pertama

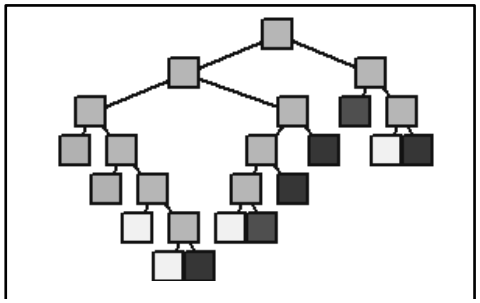
Setelah terbentuk dan terpilih kriteria pemilihan terbaik, maka simpul yang berisi 188 objek data dipilah menjadi 2 simpul. Simpul terminal 1 terbentuk akibat kriteria variabel Jurusan Biologi, Fisika, dan Matematika. Sedangkan simpul 2 terbentuk akibat kriteria variabel Jurusan Kimia sesuai dengan gambar 3.

Karena perhitungan manual cukup rumit dan membutuhkan waktu yang lama, maka digunakan bantuan *softwareSPMv8.0* untuk memudahkan peneliti dalam menentukan pemilah.

Penentuan Simpul Terminal

Tahap kedua yaitu tahap penentuan simpul terminal. Simpul *t* dikatakan sebagai simpul terminal jika tidak terdapat penurunan keheterogenan yang berarti sehingga tidak akan dipilah lagi.

Berdasarkan Gambar 4, dapat diketahui bahwa pohon klasifiksi maksimal yang terbentuk mempunyai kedalaman 7. Sedangkan simpul terminal yang dihasilkan oleh pohon klasifikasi maksimal adalah 12 simpul terminal.



Gambar 4. Pohon Klasifikasi Maksimal

Penandaan Label Kelas

Tahap ketiga adalah penandaan label kelas, berdasarkan pada persamaan (11). Sebagai contoh, pada gambar 3, untuk simpul 1.

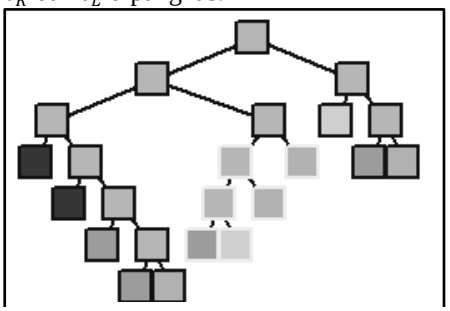
P (Lama Studi $\leq$ 5 Tahun | t)= $\frac{116}{188} = 0{,}6170$

P (Lama Studi $>$ 5 Tahun | t)= $\frac{72}{188} = 0{,}3829$

Sehingga simpul induk diberi label kelas lama studi $\leq$ 5 Tahun, karena peluang kelas tersebut lebih besar dari pada peluang kelas lainnya.

Pemangkasan Pohon Klasifikasi

Proses pemangkasan pohon klasifikasi dimulai dengan mengambil t_R yang merupakan simpul anak kanan dan t_L yang merupakan simpul anak kiri dari T_{max} yang dihasilkan dari simpul induk *t*. Jika diperoleh dua simpul anak dan simpul induk yang memenuhi persamaan (16), maka simpul anak t_R dan t_L dipangkas.



Gambar 5. Pohon Klasifikasi Maksimal yang dipangkas

Gambar 5 terdapat simpul yang akan dipangkas yaitu simpul ke 7, simpul tersebut mengalami pemangkasan karena simpul induk dan simpul anaknya memenuhi persamaan (16). Maka proses perhitungan pemangkasan adalah sebagai berikut.

Simpul induk (simpul 7) :

Dalam simpul 7 terdapat 2 kelas yaitu kelas lama studi $\leq$ 5 Tahun dan kelas lama studi > 5 Tahun. Maka nilai probabilitas tiap kelas dalam simpul 2 dengan menggunakan persamaan (11) adalah :

$$P\ (\text{Lama Studi } \leq \text{ 5 Tahun} \mid t_7) = \frac{21}{24} = 0{,}875$$

$$P\ (\text{Lama Studi } > \text{ 5 Tahun} \mid t_7) = \frac{3}{24} = 0{,}125$$

kemudian,

$$r(t_7) = 1 - 0{,}875 = 0{,}125$$

Nilai probabilitas objek yang berada dalam simpul 8 adalah :

$$P(t_7) = \frac{N(t_7)}{N} = \frac{24}{188} = 0{,}1276$$

Oleh karena itu,

$R(t_7) = r(t_7).P(t_7) = 0{,}125 \times 0{,}1276 = 0{,}01595$

Simpul anak (simpul 8) :

Simpul 8 memiliki nilai $max_j\ P(j|t)$ sebesar 0,8 , sehingga :

$$r(r_L) = 1 - max_j\ P(j|t) = 1 - 0{,}8 = 0{,}2$$

$P(t_L)$adalah peluang banyaknya objek pada simpul anak sebelah kiri, sehingga:

$$P(t_L) = \frac{N(t_L)}{N} = \frac{15}{188} = 0{,}0797$$

Oleh karena itu,

$R(t_L) = r(t_L).P(t_L) = 0{,}2 \times 0{,}0797 = 0.01595$

Simpul anak (simpul terminal 9) :

Simpul terminal 9 memiliki nilai $max_j\ P(j|t)$ sebesar 1 , sehingga :

$$r(r_R) = 1 - max_j\ P(j|t) = 1 - 1 = 0$$

$P(t_R)$adalah peluang banyaknya objek pada simpul anak sebelah kanan, sehingga:

$$P(t_R) = \frac{N(t_R)}{N} = \frac{9}{188} = 0{,}047$$

Oleh karena itu,

$R(t_R) = r(t_R).P(t_R) = 0 \times 0{,}047 = 0$

Dengan demikian, persamaan (16) :

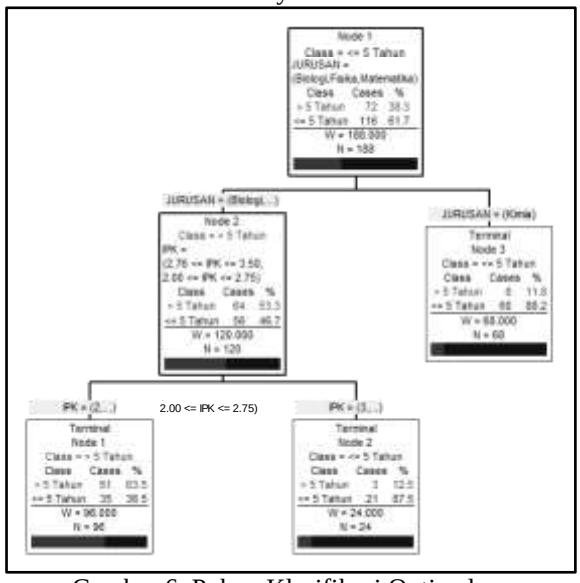
$$R(t) = R(t_L) + R(t_R)$$
$$0{,}01595 = 0{,}01595 + 0$$
$$0{,}01595 = 0{,}01595$$

Terpenuhi untuk simpul 7, sehingga dilakukan pemangkasan.

Proses ini diulangi lagi sampai tidak ada lagi pemangkasan yang mungkin terjadi, sehingga diperoleh ukuran pohon yang layak dan memenuhi *cost complexityminimum*.

Pohon Klasifikasi Optimal

Pohon klasifikasi optimal untuk data lama masa studi dapat dilihat pada Gambar 6 yang diperoleh melalui langkah pemangkasan yang telah dilakukan sebelumnya.



Gambar 6. Pohon Klasifikasi Optimal

Berdasarkan Gambar 6 dapat diketahui bahwa variabel penjelas yang menjadi pemilah utama pada pohon klasifikasi optimal adalah variabel Jurusan (X_5). Pengamatan simpul utama (simpul 1) dipilah menjadi dua simpul anak berdasarkan variabel Jurusan Biologi, Fisika dan Matematika. Dari 188 pengamatan pada simpul , sebanyak 68 pengamatan dipilah ke simpul kanan (simpul terminal 3), simpul ini tidak dipilah lagi karena telah homogen. Sedangkan sebanyak 120 pengamatan dipilah ke simpul kiri (simpul 2).

Simpul 2 dipilah lagi berdasarkan variabel pemilah IPK ($2{,}00 \leq \text{IPK} \leq 2{,}75$ dan $2{,}76 \leq \text{IPK} \leq 3{,}50$). Dari 120 pengamatan pada simpul 2, sebanyak 24 pengamatan dipilah kesimpul kanan (simpul terminal 2), sebanyak 96 pengamatan dipilah kesimpul kiri (simpul terminal 1).

Adapun interpretasi hasil untuk masing-masing simpul terminal adalah sebagai berikut :

1. Simpul Terminal 1
   Simpul terminal 1 terdiri dari 96 pengamatan yang diprediksi sebagai kelompok mahasiswa yang lama studinya $>$ 5 Tahun. Karakteristik mahasiswa dari simpul ini adalah mahasiswa yang memiliki nilai IPK $2{,}00 \leq \text{IPK} \leq 2{,}75$ dan $2{,}76 \leq \text{IPK} \leq 3{,}50$.
2. Simpul Terminal 2
   Simpul terminal 2 terdiri dari 24 pengamatan yang diprediksi sebagai kelompok mahasiswa yang lama studinya $\leq$ 5 Tahun. Karakteristik mahasiswa dari simpul ini adalah mahasiswa yang memiliki nilai IPK $3.51 \leq \text{IPK} \leq 4.00$
3. Simpul Terminal 3
   Simpul terminal 3 terdiri dari 68 pengamatan yang diprediksi sebagai kelompok mahasiswa yang lama studinya $\leq$ 5 Tahun. Karakteristik mahasiswa dari simpul ini adalah mahasiswa yang berasal dari jurusan Kimia.

Ketepatan Klasifikasi

Pohon klasifikasi optimal yang telah terpilih tadi kemudian diuji tingkat keakuratannya dalam mengelompokkan data, yaitu diperoleh ketepatan pengklasifikasi sebesar :

$$\frac{61 + 81}{188} \times 100\% = 77{,}27\ \%$$

Berdasarkan perhitungan ketepatan klasifikasi diperoleh 77,27 % yang artinya pohon klasifikasi yang terbentuk mampu memprediksi dengan tepat pengamatan sebesar 77,27 %.

Kesimpulan

Berdasarkan hasil analisis dan pembahasan yang dilakukan, kesimpulan yang diperoleh dari penelitian ini, yaitu :

1. Berdasarkan hasil penelitian menggunakan metode Regresi Logistik Biner, Faktor – faktor yang mempengaruhi Lama Studi Mahasiswa FMIPA Universitas Mulawarman adalah IPK , Jenis Kelamin, Jenis Sekolah Menengah dan Jurusan .
2. Berdasarkan hasil penelitian menggunakan metode CART, Hasil pengklasifikasian yang diperoleh adalah mahasiswa yang lama studinya $\leq$ 5 Tahun adalah mahasiswa yang berasal dari jurusan Kimia atau memiliki IPK $3.51 \leq \text{IPK} \leq 4.00$. Sedangkan mahasiswa yang lama studinya > 5 Tahun yaitu mahasiswa yang memiliki IPK $2.00 \leq \text{IPK} \leq 2.75$; $2.76 \leq \text{IPK} \leq 3.50$.
3. Pada kasus Lama Studi Mahasiswa FMIPA Universitas Mulawarman kinerja CART lebih baik dibandingkan dengan Regresi Logistik Biner. Hal ini dapat dilihat dari ketepatan klasifikasi CART sebesar 77,27 % sedangkan ketepatan klasifikasi Regresi Logistik Biner 75 %.

Daftar Pustaka

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

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression.* New York : John Wiley & Sons, Inc.

Johnson, R.A & DW Wichen, (2007). *Applied Multivariat Statistical Sixth Edition* , Prentice Hall International Inc, New Jersey.

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

Timofeev, Roman. (2004). *Classification and Regression Tree (CART) Theory and Applications.* A Master Tesis. CASE- Center of Applied Statistics and Economics. Berlin : Humboldt University.