IDENTIFIKASI FAKTOR-FAKTOR YANG MEMPENGARUHI MAHASISWA PASCASARJANA IPB BERHENTI STUDI MENGGUNAKAN ANALISIS CHAID DAN REGRESI LOGISTIK

Mohamad Jajuli

Teknik Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
mohamad.jajuli@staff.unsika.ac.id

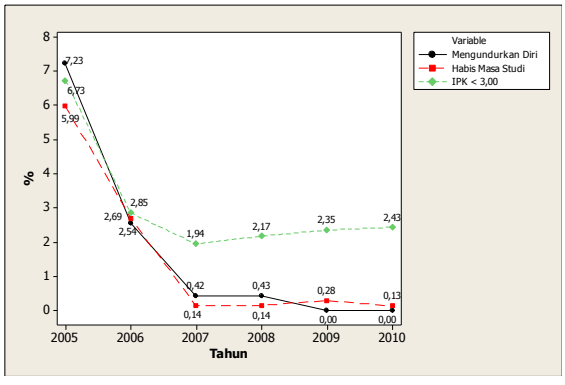
ABSTRAK

Institut Pertanian Bogor (IPB) berusaha semaksimal mungkin meningkatkan kelulusan para mahasiswanya baik secara kualitas maupun kuantitas. Secara kualitas mahasiswa lulus dengan nilai IPK yang maksimal dan lulus tepat waktu. Secara kuantitas artinya jumlah mahasiswa yang masuk sama dengan jumlah mahasiswa yang lulus, berarti tidak ada yang berhenti studi. Mahasiswa berhenti studi merupakan salah satu persoalan yang dapat merugikan pribadi mahasiswa, institusi, dan negara. Keterkaitan mahasiswa pascasarjana IPB berhenti studi berdasarkan jenis kelamin, usia, status perkawinan, status pekerjaan, status perguruan tinggi asal, IPK S1, sumber biaya S2, daerah perguruan tinggi asal, dan linieritas S1 dapat dilihat dengan analisis *Chisquared Automatic Interaction Detection* (CHAID) dan regresi logistik. Data yang digunakan dalam penelitian ini adalah data mahasiswa pascasarjana IPB angkatan 2005-2010. Berdasarkan analisis CHAID menghasilkan 4 faktor yang mempengaruhi mahasiswa pascasarjana IPB berhentis studi yaitu sumber biaya pendidikan S2, status perguruan tinggi asal, linieritas S1, dan IPK S1. Hasil analisis regresi logistik menunjukkan bahwa mahasiswa pascasarjana IPB berhenti studi dipengaruhi oleh jenis kelamin, status perguruan tinggi asal, sumber biaya pendidikan S2, dan linearitas rumpun ilmu. Ketepatan klasifikasi yang diperoleh untuk analisis CHAID dan regresi logistik masing-masing sebesar 97.1%.

Keywords: berhenti studi, CHAID, regresi logistik

PENDAHULUAN

Sekolah Pascasarjana Institut Pertanian Bogor (IPB) dengan mottonya meraih masa depan berkualitas bersama pascasarjana IPB, berusaha semaksimal mungkin meningkatkan kelulusan para mahasiswanya baik secara kualitas maupun kuantitas. Secara kualitas mahasiswa lulus dengan nilai indeks prestasi kumulatif (IPK) yang maksimal dan lulus tepat waktu. Secara kuantitas artinya jumlah mahasiswa yang masuk/terdaftar sama dengan jumlah mahasiswa yang lulus, berarti tidak ada yang berhenti studi. Kategori mahasiswa berhenti studi di sekolah pascasarjana IPB sendiri ada tiga, yaitu mahasiswa mengundurkan diri, habis masa studi, dan IPK di bawah 3,00.



Gambar 1. Persentase kategori berhenti studi mahasiswa pascasarjana IPB

Pada Gambar 1 terlihat bahwa mahasiswa mengundurkan diri dan habis masa studi persentasenya mengalami penurunan di tahun 2006-2010, sedangkan mahasiswa yang IPK kurang dari 3,00 mengalami penurunan di tahun 2006 tetapi perlahan meningkat kembali di tahun 2008 hingga 2010. Dilihat dari total mahasiswa berhenti studi, persentase mahasiswa yang IPK kurang dari 3,00 sebesar 52,88% sedangkan mahasiswa mengundurkan diri sebesar 24,52% dan mahasiswa yang habis masa studi 22,60%. Hal ini membuktikan bahwa sebagian besar mahasiswa pascasarjana IPB berhenti studi disebabkan karena IPK yang kurang dari 3,00 sehingga dalam penelitian ini yang akan dikaji lebih lanjut adalah mahasiswa pascasarjana IPB yang berhenti studi karena IPK kurang dari 3,00.

Metode statistika yang dapat melihat karakteristik mahasiswa yang berhenti studi dan tidak berhenti studi yaitu dengan menggunakan CHAID (*Chi-squared Automatic Interaction Detection*) dan regresi logistik. CHAID merupakan salah satu teknik nonparametrik yang dapat melakukan pemilihan variabel dari data berukuran besar dalam menentukan variable-variabel yang paling berpengaruh. CHAID akan menghasilkan diagram yang mirip dengan diagram pohon keputusan dan menggunakan uji Khi-kuadrat pada pengoperasiannya. Metode ini cocok digunakan pada data yang berukuran besar dan akan menghasilkan pohon nonbiner (Alamudi *et al.*1998).

Regresi logistik adalah suatu teknik analisis statistika yang digunakan untuk mendeskripsikan hubungan antara variabel respon yang memiliki dua kategori atau lebih dengan satu atau lebih variabel penjelas berskala kategori atau kontinu (Hosmer dan Lemeshow 2000).

Tujuan penelitian ini adalah menentukan factor-faktor yang mempengaruhi mahasiswa pascasarjana IPB berhenti studi menggunakan analisis CHAID dan regresi logistik.

TINJAUAN PUSTAKA

CHAID (*Chi-squared Automatic Interaction Detection*)

CHAID adalah singkatan dari *Chi-squared Automatic Interaction Detector*. CHAID pertama kali diperkenalkan dalam sebuah artikel berjudul "*An Exploratory Technique for Investigating Large Quantities of Categorical Data*" oleh Dr. G.V. Kass tahun 1980. Prosedurnya merupakan bagian dari teknik terdahulu yang dikenal dengan *Automatic Interaction Detector* (AID), dan menggunakan statistik *chi-square* sebagai alat utamanya. CHAID secara keseluruhan bekerja untuk menduga sebuah variabel tunggal, disebut sebagai variabel dependen, yang didasarkan pada sejumlah variabel-variabel yang lain, disebut sebagai variabelvariabel independen. CHAID merupakan suatu teknik iteratif yang menguji satu-persatu variabel independen yang digunakan dalam klasifikasi, dan menyusunnya berdasarkan pada tingkat signifikansi statistik *chisquare* terhadap variabel dependennya (Gallagher, 2000).

CHAID digunakan untuk membentuk segmentasi yang membagi sebuah sampel menjadi dua atau lebih kelompok yang berbeda berdasarkan sebuah kriteria tertentu. Hal ini kemudian diteruskan dengan membagi kelompok-kelompok tersebut menjadi kelompok yang lebih kecil berdasarkan variable variabel independen yang lain. Prosesnya berlanjut sampai tidak ditemukan lagi variabel independen yang signifikan secara statistik. Segmen-segmen yang dihasilkan akan bersifat saling lepas yang secara statistik akan memenuhi criteria pokok segmentasi dasar (Bagozzi, 1994). Hasilnya juga akan memberikan peringkat pada variabel yang merupakan variabel independen paling signifikan sampai yang tidak signifikan.

CHAID memilih variabel-variabel variable independennya atas dasar uji *chi-square* antara kategori variabel-variabel yang tersedia dengan kategori-kategori variabel dependennya (seperti yang terdapat pada statistika dasar bahwa uji *chi-square* merupakan uji non parametrik yang sesuai untuk menguji hubungan antar variabel yang berbentuk kategori) (Myers, 1996).

Regresi Logistik

Hosmer dan Lemeshow (2000) menjelaskan bahwa model regresi logistik dibentuk dengan menyatakan nilai $P(Y=1|x)$ sebagai $\pi(x)$, yang dinotasikan sebagai berikut:

$$\pi(x)=\frac{\exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}{1+\exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}$$

Fungsi regresi di atas berbentuk non linear sehingga untuk membuatnya menjadi fungsi linear dilakukan transformasi logit sebagai berikut:

$$\text{logit}[\pi(x)]=\beta_0+\beta_1X_1+...+\beta_pX_p=\ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

Untuk variabel bebas yang bersifat kategorik maka diperlukan variabel boneka. Secara umum jika sebuah variabel berskala nominal atau ordinal mempunyai k kategori, maka diperlukan k-1 variabel boneka. Misalnya, variabel penjelas ke-j mempunyai k_j kategori. D_{ju} melambangkan k_j-1 variabel boneka dan β_{ju} merupakan koefisien variabel boneka dengan $u=1,2,...,k_j-1$. Dengan demikian model transformasi logitnya menjadi:

$$g(x)=\beta_0+\beta_1x_1+...+\sum_{u=1}^{k_j-1}\beta_{ju}D_{ju}+\beta_px_p$$

Pendugaan parameter dalam model regresi logistik dilakukan dengan menggunakan metode kemungkinan maksimum. Jika antara amatan yang satu dengan yang lain diasumsikan bebas, maka fungsi kemungkinannya adalah:

$$l(\beta)=\prod_{i=1}^{n}\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}$$

dengan:

$i = 1, 2, ..., p$

y_i = pengamatan pada variabel respon ke-i

$\pi(x_i)$ = peluang untuk variabel penjelas ke-i bernilai Y=1

Koefisien logit diduga dengan memaksimumkan $l(\beta)$ dengan pendekatan logaritma sehingga fungsinya sebagai berikut:

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

Nilai dugaan β_i dapat diperoleh dengan membuat turunan pertama $L(\beta)$ terhadap $\beta_i = 0$ dengan $i = 1, 2, ..., p$.

Menguji peranan dari tiap variabel penjelas terhadap variabel responnya dalam regresi logistik menggunakan statistik uji G dan uji Wald. Statistik uji G adalah uji rasio kemungkinan yang digunakan untuk menguji peranan variabel penjelas di dalam model secara serentak. Hipotesis yang diuji yaitu:

H_0: $\beta_1 = ... = \beta_p = 0$

H_1: minimal ada satu $\beta_i \neq 0$; $i = 1, 2, ..., p$

Rumus umum untuk uji G adalah:

$$G=-2\ln\left[\frac{L_0}{L_p}\right]$$

dengan L_0 adalah fungsi kemungkinan tanpa variabel penjelas dan L_p merupakan fungsi

kemungkinan dengan p variabel penjelas. Hipotesis nol ditolak jika $G > \chi^2_{p(\alpha)}$.

Statistik uji Wald digunakan untuk menguji parameter β_i secara parsial. Hipotesis yang akan diuji adalah:

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$; i = 1, 2, ..., p

Statistik uji Wald adalah:

$$W_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}i)}$$

dengan $\hat{\beta}_i$ sebagai penduga β_i dan $SE(\hat{\beta}_i)$ sebagai penduga galat baku β_i. Hipotesis nol ditolak jika $|W| > Z_{\alpha/2}$.

Interpretasi koefisien untuk model regresi logistik biner dapat dilakukan dengan melihat nilai rasio oddsnya. Rasio odds merupakan ukuran asosiasi yang memperkirakan berapa besar kecenderungan pengaruh variabel-variabel penjelas terhadap variabel respon. Rasio odds (ψ) didefinisikan sebagai berikut

$$\hat{\Psi} = \exp(\beta_i)$$

METODOLOGI PENELITIAN

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari bagian akademik pascasarjana Institut Pertanian Bogor (IPB) yaitu data mahasiswa pascasarjana IPB angkatan 2005-2010. Variabel yang digunakan dalam penelitian ini sebagai berikut:

Variabel dependen

Y = Status Mahasiswa (tidak berhenti studi, berhenti studi)

Variabel independen

X_1 = Jenis kelamin (perempuan, laki-laki)

X_2 = Usia (< 33 tahun,33-49 tahun, > 49 tahun)

X_3 = Status Perkawinan (lajang, menikah)

X_4 = Status pekerjaan (tidak bekerja, bekerja)

X_5 = Status perguruan tinggi asal (swasta, negeri)

X_6 = IPK S1 (< 2,75, ≥ 2,75)

X_7 = Sumber biaya S2 (mandiri, beasiswa)

X_8 = Daerah perguruan tinggi asal (pulau Jawa, luar pulau Jawa)

X_9 = Linieritas S1 (tidak linier, linier)

Prosedur Analisis CHAID

Langkah-langkah dalam melakukan analisis CHAID secara garis besar adalah sebagai berikut:

1. Memasukkan semua data berdasarkan kategori yang ditentukan
2. Menentukan terlebih dahulu semua skala variabel yang akan digunakan dengan tepat dan benar
3. Menentukan kategori target dari kategori-kategori variabel dependen. Hal ini dilakukan untuk memunculkan beberapa grafik lain sebagai informasi lebih lanjut dalam data yang ada. Kategori target yang dipergunakan bisa salah satu atau semua kategori yang ada pada variabel dependen.
4. Menerapkan 3 langkah analisis CHAID, yaitu langkah penggabungan, pemisahan, dan pemberhentian. Dalam langkah penggabungan akan mulai diterapkan uji *Chi-square* dan pengali Bonferroni sebagai pengoreksinya. Pada langkah penggabungan sebagian besar proses akan menggunakan uji *Chi-square* saja. Kemudian dilakukan iterasi pada kedua langkah tersebut, dan proses iterasi akan berhenti apabila sudah tidak ada lagi variabel independen yang tersisa untuk diuji hubungannya dengan variabel dependen, atau juga apabila terbentuknya node pada diagram pohon telah memenuhi batasan yang ditentukan. Proses ini disebut dengan proses pemberhentian
5. Menentukan kelompok faktor yang mempengaruhi mahasiswa berhenti studi dengan menginterpretasikan diagram pohon CHAID

Prosedur Analisis Regresi Logistik

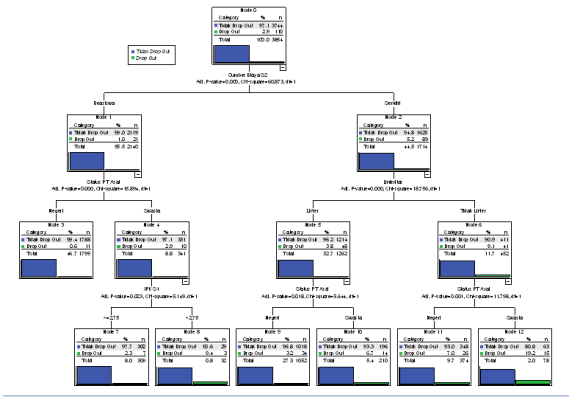
Langkah-langkah dalam melakukan analisis regresi logistik secara garis besar adalah sebagai berikut:

1. Menduga parameter regresi logistic
2. Menguji kontribusi variable secara simultan
3. Menguji kontribusi variable secara parsial
4. Menguji kebaikan model regresi logistik

HASIL DAN PEMBAHASAN

Analisis CHAID

Hasil analisis CHAID menunjukkan bahwa hanya ada 4 dari 9 faktor yang mempengaruhi mahasiswa pascasarjana IPB berhenti studi, yaitu variabel sumber biaya S2, linieritas S1, status perguruan tinggi asal, dan IPK S1.



Gambar 2. Dendogram hasil analisis CHAID

Gambar 2 menunjukkan bahwa mahasiswa yang sumber biaya S2 nya mandiri cenderung drop out dibandingkan yang beasiswa. Mahasiswa yang sumber biaya S2 nya beasiswa dan status perguruan tinggi asal swasta cenderung drop out dibandingkan mahasiswa yang sumber biaya S2 nya beasiswa dan status perguruan tinggi asal negeri. Mahasiswa yang

sumber biaya S2 nya beasiswa dan status perguruan tinggi asal swasta serta IPK S1 < 2.75 cenderung drop out dibandingkan Mahasiswa yang sumber biaya S2 nya beasiswa dan status perguruan tinggi asal swasta serta IPK S1 ≥ 2.75.

Dari Gambar 2 juga terlihat bahwa mahasiswa yang sumber biaya S2 nya mandiri dan S1 nya tidak linier dengan S2 yang diambil cenderung drop out dibandingkan mahasiswa yang sumber biaya S2 nya mandiri dan S1 nya linier. Mahasiswa yang sumber biaya S2 nya mandiri dan S1 nya linier dengan S2 yang diambil serta status perguruan tinggi asalnya swasta cenderung drop out dibandingkan mahasiswa yang sumber biaya S2 nya mandiri dan S1 nya linier serta status perguruan tinggi asalnya negeri. Mahasiswa yang sumber biaya S2 nya mandiri dan S1 nya tidak linier dengan S2 yang diambil serta status perguruan tinggi asalnya swasta cenderung drop out dibandingkan mahasiswa yang sumber biaya S2 nya mandiri dan S1 nya tidak linier serta status perguruan tinggi asalnya negeri.

Regresi Logistik

Faktor apa saja yang mempengaruhi mahasiswa pascasarjana IPB berhenti studi angkatan 2005-2010 dilihat berdasarkan status mahasiswa berhenti studi atau tidak dengan menggunakan analisis regresi logistik biner.

Tabel 1. Analisis regresi logistik Biner

Variabel	Dugaan	Uji Wald	Nilai-p	Odds Ratio
Intersep	-2,122	22,650	**0,000**	-
Jenis Kelamin	0,506	5,950	**0,015**	1,66
Usia 1	0,480	2,801	0,094	1,62
Usia 2	0,685	1,409	0,235	1,98
Status Perkawinan	-0,035	0,022	0,881	0,97
Status Pekerjaan	-0,317	2,245	0,134	0,73
Status PT Asal	-0,999	21,699	**0,000**	0,37
IPK S1	-0,130	0,202	0,653	0,88
Sumber Biaya S2	-1,715	41,038	**0,000**	0,18
Daerah PT Asal	0,265	1,580	0,209	1,30
Linieritas S1	-0,737	13,295	**0,000**	0,48

Model logit untuk faktor-faktor berhenti studi mahasiswa pascasarjana IPB angkatan 2005-2010 adalah sebagai berikut:

$$\hat{g}(x)=-2,122+0,506X_1+0,480X_2(1)$$
$$+ 0,685X_2(2)-0,035X_3-0,317X_4-0,999X_5$$
$$-0,130X_6-1,715X_7+0,265X_8-0,737X_9$$

Tabel 2. Uji parameter model logit

Variabel	G hitung	p-value
Intersep		
Jenis Kelamin		
Usia 1		
Usia 2		
Status Perkawinan		
Status Pekerjaan	116,487	0,000
Status PT Asal		
IPK S1		
Sumber Biaya S2		
Daerah PT Asal		
Linieritas S1		

Pada Tabel 2 didapatkan uji G bernilai 116,487 dengan nilai p-value sebesar $0,000<\alpha$ $(0,05)$ maka paling tidak ada satu variabel independen yang berpengaruh nyata terhadap variabel dependen. Pada uji Wald di Tabel 1 didapat bahwa ada empat variabel independen yang berpengaruh nyata terhadap dependen yaitu jenis kelamin, status PT asal, sumber biaya pendidikan, dan linieritas S1 yang berarti bahwa mahasiswa pascasarjana IPB berhenti studi dipengaruhi oleh jenis kelamin, status PT asal, sumber biaya pendidikan, dan linieritas S1 dari si mahasiswa tersebut.

Ketepatan Klasifikasi Analisis CHAID dan Regresi Logistik

Kebaikan model dapat dilihat melalui tabel ketepatan klasifikasi. Berdasarkan Tabel 3, analisis CHAID menghasilkan 97.1% dari 3854 mahasiswa yang diprediksi tepat.

Tabel 3. Ketepatan klasifikasi hasil analisis CHAID

Aktual	Prediksi		
	Tidak Berhenti Studi	Berhenti Studi	% Ketepatan
Tidak Berhenti Studi	3744	0	100%
Berhenti Studi	110	0	0%
Persentase Keseluruhan	100%	0%	97,1%

Tabel 4 menunjukkan ketepatan klasifikasi analisis regresi logistik yang hasilnya sama dengan ketepatan klasifikasi analisis CHAID yaitu menghasilkan 97.1% dari 3854 mahasiswa yang diprediksi tepat.

Tabel 4. Ketepatan klasifikasi hasil analisis regresi logistik

Aktual	Prediksi		
	Tidak Berhenti Studi	Berhenti Studi	% Ketepatan
Tidak Berhenti Studi	3744	0	100%
Berhenti Studi	110	0	0%
% Ketepatan Keseluruhan (CCR)			97,1%

KESIMPULAN

Hasil dari analisis CHAID, variabel independen yang berhubungan dengan mahasiswa pascasarjana IPB berhenti studi adalah variabel sumber biaya S2, linieritas S1, status perguruan tinggi asal, dan IPK S1 dan menghasilkan kriteria mengenai mahasiswa pascasarjana IPB berhenti studi. Kriteria pertama adalah mahasiswa yang sumber biaya S2 nya mandiri, kriteria kedua mahasiswa yang sumber biaya S2 nya beasiswa dan status perguruan tinggi asal swasta, kriteria ketiga mahasiswa yang sumber biaya S2 nya beasiswa dan status perguruan tinggi asal swasta serta IPK S1 < 2.75, kriteria keempat mahasiswa yang sumber biaya S2 nya mandiri dan S1 nya tidak linier dengan S2 yang diambil, kriteria kelima mahasiswa yang sumber biaya S2 nya mandiri dan S1 nya linier dengan S2 yang diambil serta status perguruan tinggi asalnya swasta, kriteria keenam mahasiswa yang sumber biaya S2 nya mandiri dan S1 nya tidak linier dengan S2 yang diambil serta status perguruan tinggi asalnya swasta.

Berdasarkan hasil analisis regresi logistik bahwa mahasiswa pascasarjana IPB berhenti studi dipengaruhi oleh jenis kelamin, status PT asal, sumber biaya pendidikan S2, dan linieritas rumpun ilmu.

Ketepatan klasifikasi yang diperoleh untuk analisis CHAID dan regresi logistik memiliki nilai yang sama yaitu sebesar 97.1%.

SARAN

Penelitian selanjutnya diharapkan mengkaji faktor-faktor internal seperti kondisi fisiologis, psikologis, panca indera, intelegensi, bakat, dan motivasi dari mahasiswa pascasarjana IPB.

DAFTAR PUSTAKA

Alamudi A, Wigena AH, Aunuddin. 1998. Eksplorasi struktur data dengan metode CHAID. Forum Statistika dan Komputasi 3(1):10-16.

Agresti, A. 2007. *An* Introduction to *Categorical Data Analysis*. New York: John Wiley & Sons.

Antipov E, Pokryshevskaya E. 2010. *Applying CHAID for logistic regression diagnostics and classification accuracy improvement*. Journal of Targeting, Measurement and Analysis for Marketing 18(2):109–117.

Breiman L, JH Friedman, RA Olshen, CJ Stone. 1993. *Classification and Regression Trees*. New York (US): Champan and Hall.

Cameron, A. C., Trivedi, P. K. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Hardin, J. W., Hilbe, J. M. 2007. *Generalized Linear Models and Extensions*. Texas: A Stata Press Publication.

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

Khosgoftaar, T. M., Gao, K., Szabo, R. M. 2004. *Comparing Software Fault Predictions of Pure and Zero-Inflated Poisson Regression Models*. International Journal of System Science 36(11):705-715.

McCullagh, P., J. A. Nelder. 1983. *Generalized Linear Models*. London: Chapmann and Hall.

Myers, R. H. 1990. *Classical and Modern Regression with Applications*. Ed ke-2. PWS-KENT Publishing Company. Boston.