KLASIFIKASI DENGAN METODE CHAID
(*CHI-SQUARED AUTOMATIC INTERACTION DETECTION*)
DAN PENERAPANNYA PADA KLASIFIKASI ALUMNI FMIPA UNY

SKRIPSI

Diajukan Kepada Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Negeri Yogyakarta
Untuk Memenuhi Sebagian Persyaratan
Guna Memperoleh Gelar Sarjana Sains

Diajukan oleh:

Husein Permana
06305141032

PROGRAM STUDI MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI YOGYAKARTA
2011

PENGESAHAN

Skripsi

Klasifikasi dengan Metode CHAID

(*Chi-Squared Automatic Interaction Detectin*)

dan Penerapannya pada Klasifikasi Alumni FMIPA UNY

Disusun Oleh :

Husein Permana

06305141032

Telah Dipertahankan Di Depan Panitia Penguji Skripsi Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Yogyakarta pada tanggal 24 Januari 2011 dan dinyatakan telah memenuhi syarat guna memperoleh gelar sarjana sains.

Susunan Panitia Penguji Skripsi

Nama	Jabatan	Tanda Tangan	Tanggal
Dr. Dhoriva U.W.	Ketua Penguji	…….	…….
Retno Subekti, M.Sc	Sekretasis Penguji	…….	…….
Dr. Djamilah	Penguji Utama	…….	…….
Mathilda S, M.Si	Penguji Pendamping	…….	…….

Yogyakarta, Januari 2011

Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Negeri Yogyakarta

Dekan

Dr. Ariswan

NIP. 195909141988031003

PERSETUJUAN

SKRIPSI

KLASIFIKASI DENGAN METODE CHAID

(*CHI-SQUARED AUTOMATIC INTERACTION DETECTION*)

DAN PENERAPANNYA PADA KLASIFIKASI ALUMNI FMIPA UNY

Oleh:

Husein Permana

06305141032

Telah Disetujui pada Tanggal 17 Januari 2011

Untuk Dipertahankan di Depan Panitia Penguji Tugas Akhir Skripsi

Program Studi Matematika

Jurusan Pendidikan Matematika

Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Negeri Yogyakarta

Menyetujui,

Dosen Pembimbing

Dr. Dhoriva U.W

NIP. 196603311993032001

SURAT PERNYATAAN

Dengan ini saya menyatakan bahwa skripsi ini benar-benar karya saya sendiri. Sepanjang pengetahuan saya tidak terdapat karya atau pendapat yang ditulis atau diterbitkan orang lain kecuali sebagai acuan atau kutipan dengan mengikuti tata penulisan karya ilmiah yang telah lazim.

Yogyakarta, 17 Januari 2011
Yang Menyatakan,

Husein Permana

HALAMAN PERSEMBAHAN

Kupersembahkan karya kecil ini dengan kasih sayang untuk :

Kedua orang tuaku,

Almarhumah Ibu yang selalu kurindukan dan

Bapak sebagai ayah yang sangat hebat

serta

mbak Nanu & de' Imas

Ucapan Terima kasih untuk :

1. Allah SWT, yang senantiasa mencurahkan Rahmat dan HidayahNya
2. Nabi Muhammad Saw yang telah memberikan pencerahan kepada umat manusia dan suri tauladan yang baik.
3. Ibu Dr. Dhoriva U.W, selaku dosen pembimbing yang telah memberikan arahan dan bimbingan Tugas Akhir Skripsi. *Matursembahnuwun....*
4. Mbak Nanu, terimakasih buat masukan-masukannya dalam penulisan skripsi ini.
5. Farah, terimakasih buat dukungan & perhatian yang tak pernah habis, Gumawo...
6. Teman-temanku warga Matematika Reg'06 khususnya Puguh, Ginanjar, Hermawan, Eko, Adit Jan's, Kholis, Ifa dan Deeya.
7. Semua OP LIMUNY PUSKOM UNY, terimakasih untuk kebersamaan, kekompakan dan kekeluargaan yang kalian berikan. Selalu merindukan masa-masa kerja banting tulang siang dan malam demi LIMUNY tercinta. Jayalah LIMUNY..!!!
8. Sahabat-sahabatku sejak masa SMA, Susilo, Tertian, Ninis, Ati. Ayok tetap semangat mengejar masa depan! Fight! Fight! Fight!

MOTTO

"Allah mengangkat orang-orang yang beriman dari golonganmu dan juga orang-orang yang dikaruniai ilmu pengetahuan hingga beberapa derajat"
(Q.S. Al-Mujaadilah; 58:11)

"Kegigihan adalah semangat pantang menyerah yang harus kita miliki. Dengan bekal kegigihan dan usaha yang konsisten, kesuksesan yang kita peroleh pasti berkualitas dan membanggakan"
(Andrie Wongso)

"5 S : Senyum, Salam, Sapa, Sopan, dan Santun, 3 M : Mulailah dari diri sendiri, Mulailah dari hal yang kecil, Mulailah saat ini"
(KH. Abdullah Gymnastiar)

Optimisme adalah suatu keharusan dalam segala hal, dengan optimis maka kita akan mempunyai kekuatan untuk mendapatkan hasil terbaik karena setelah kesulitan akan datang kemudahan.

Klasifikasi dengan Metode CHAID
(Chi-Squared Automatic Interaction Detection)
dan penerapannya pada Klasifikasi Alumni S1 FMIPA UNY

Disusun Oleh :
Husein Permana
06305141032

ABSTRAK

Klasfikasi merupakan proses mengelompokkan suatu data menjadi kelompok-kelompok yang lebih kecil dengan ciri yang relatif sama. Salah satu metode klasifikasi adalah metode CHAID (*Chi-Squared Automatic Interaction Detection*). Metode CHAID akan membagi data menjadi beberapa kelompok/segmen yang lebih kecil dari data berdasarkan hubungan variabel dependen dengan variabel independen. Penyusunan Skripsi ini bertujuan untuk menjelaskan prosedur klasifikasi dengan metode CHAID, menerapkan metode CHAID pada kasus klasifikasi alumni FMIPA UNY berdasarkan masa studinya dan mengetahui perbandingan hasil klasifikasi metode CHAID dengan analisis regresi logistik biner.

Prosedur klasifikasi dengan metode CHAID dapat dinyatakan dengan 3 langkah pokok, yaitu penggabungan (*merging*), pemisahan (*splitting*) dan penghentian (*stopping*). Penerapan metode CHAID menghasilkan 4 variabel independen yang signifikan terhadap model, yaitu IP semester 1, program studi, jalur masuk dan asal daerah. Analisis regresi logistik biner hanya menghasilkan 2 variabel independen yang signifikan terhadap model yaitu IP semester 1 dan program studi. Metode CHAID menyebutkan bahwa segmen alumni yang memiliki persentase masa studi tidak tepat waktu paling besar adalah alumni yang IP semester 1 kurang dari sama dengan 2,50, program studi non kependidikan dan berasal dari jalur masuk non regular. Segmen alumni yang memiliki persentase lulus tepat waktu paling besar adalah alumni yang IP semester 1 lebih dari sama dengan 3,51. Dalam kasus ini akurasi hasil klasifikasi kedua metode tersebut tidak jauh berbeda, namun metode CHAID dianggap lebih unggul karena dapat membagi alumni membagi beberapa segmen, sedangkan regresi logistik biner hanya mengklasifikasikan alumni menjadi kategori lulus tepat waktu dan tidak tepat waktu. Pembagian segmen-segmen ini mempermudah penindaklanjutan hasil klasifikasi karena bisa diketahui dengan tepat segmen mana yang menjadi prioritas.

KATA PENGANTAR

Alhamdulillahirobbil'alamiin, segala puji bagi Allah SWT yang telah mencurahkan rahmat dan Karunia-Nya sehingga penulis mampu menyelesaikan penulisan Skripsi dengan judul "**Klasifikasi dengan Metode CHAID (Chi-Squared Automatic Interaction Detection) dan penerapannya pada Klasifikasi Alumni FMIPA UNY"** ini dengan baik. Penulisan Skripsi ini disusun dalam rangka memenuhi persyaratan untuk memperoleh gelar Sarjana Sains Program Studi Matematika di Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Yogyakarta.

Penulis menyadari sepenuhnya bahwa dalam penulisan skripsi ini tidak terlepas dari dukungan, motivasi, kerjasama maupun bimbingan dari berbagai pihak. Oleh karena itu, penulis mengucapkan terimakasih yang sebesar-besarnya kepada :

1. Bapak Dr. Ariswan, selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Yogyakarta yang telah memberikan kesempatan penulis dalam menyelesaikan studi.
2. Bapak Dr. Hartono, Ketua Jurusan Pendidikan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Yogyakarta yang telah memberikan kemudahan pengurusan administrasi.
3. Ibu Atmini Dhoruri, M.Si, Ketua Program Studi Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Yogyakarta yang telah memberi dukungan untuk kelancaran studi.

4. Ibu Dr. Dhoriva U.W. selaku dosen pembimbing yang telah dengan sabar membimbing penulis dan selalu memberikan motivasi kepada penulis.
5. Bapak Fauzan M.Si, dosen penasehat akademik penulis.
6. Semua pihak yang telah membantu tersusunnya skripsi ini yang tidak dapat penulis sebutkan satu-persatu.

Penulis menyadari bahwa dalam skripsi ini masih banyak sekali kekurangan dan kesalahan. Oleh karena itu penulis mengharapkan kritik dan saran yang membangun untuk menyempurnakan skripsi ini. Akhir kata, penulis berharap semoga skripsi ini dapat memberikan sesuatu yang bermanfaat bagi semua pihak yang membacanya.

Yogyakarta, Januari 2011

Penulis

DAFTAR ISI

Abstrak ……... viii

Kata Pengantar ……... ix

Daftar Isi …… xi

Daftar Tabel ……... xiii

Daftar Gambar ……... xiv

Daftar lampiran …….. xv

BAB I PENDAHULUAN

A. Latar Belakang Masalah …… 1

B. Rumusan Masalah …….. 4

C. Tujuan Penulisan …… 4

D. Manfaat Penulisan …….. 5

BAB II Kajian Pustaka

A. Klasifikasi …….. 6

B. Uji Independensi *Chi-Square* (χ^2) …….. 7

C. Regresi Nonlinier ……... 10

D. *Linier Probability Model* (LPM) dan Transformasi Logit ……. 11

E. Metode Maksimum *Likelihood* …….. 15

F. Metode Neton-Raphson ……. 18

G. Inferensi dalam Regresi Logistik ……... 19

H. Peluang Kejadian ……... 21

I. Evaluasi Kasil Klasifikasi …….. 21

BAB III PEMBAHASAN

A. Analisis CHAID ……. 24

1. Variabel-Variabel dalam Analisis CHAID …….. 25

2. Algoritma CHAID ……... 26

i. Penggabungan (*Merging*) …… 26

ii. Pemisahan (*Splitting*) ……... 29

iii. Penghentian (*Stopping*) …… 30

3. Koreksi Bonferoni (*Bonferroni Correction*) …… 31

4. Diagram Pohon Klasifikasi CHAID …… 33

B. Klasifikasi Alumni FMIPA UNY …….. 34

1. Deskripsi Variabel ……... 37

2. Analisis Data dengan Metode CHAID …… 38

3. Analisis Data dengan Regresi Logistik Biner …….. 46

C. Perbandingan Hasil Klasifikasi Metode CHAID dan Regresi Logistik biner ……. 52

BAB IV KESIMPULAN DAN SARAN

A. Kesimpulan …… 54

B. Saran …….. 57

DAFTAR PUSTAKA …… 58

LAMPIRAN ……... 59

DAFTAR TABEL

Tabel 2.1 Struktur data uji *Chi Square* 8

Tabel 2.2 Peluang kejadian 9

Tabel 2.3 *Confusius matrix* 22

Tabel 3.1 Ilustrasi pasangan penggabungan variabel 28

Tabel 3.2 Pengkategorian alumni dan pemberian kode berdasarkan asal daerahnya 35

Tabel 3.3 Pengkategorian alumni dan pemberian kode berdasarkan IP semester 1 37

Tabel 3.4 Ringkasan pembentukan model dengan metode CHAID 41

Tabel 3.5 Segmentasi alumni FMIPA 44

Tabel 3.6 Persentase setiap segmen alumni FMIPA UNY 45

Tabel 3.7 Klasifikasi dari metode CHAID 46

Tabel 3.8 Pembentukan variabel semu 48

Tabel 3.9 *Variables in the equation* 49

Tabel 3.10 Omnibus test of model coeffisients ……..... 50

Tabel 3.11 Klasifikasi analisis regresi logistik biner …… 51

Tabel 4.1 Perbandingan hasil klasifikasi metode CHAID dengan regresi logistik biner …… 55

DAFTAR GAMBAR

Gambar 2.1 Contoh *decision tree* pada klasifikasi penyakit hipertensi 6

Gambar 2.2 Kurva regresi logistik 14

Gambar 3.1 Diagram alir algoritma CHAID 31

Gambar 3.2 Diagram pohon dalam analisis CHAID 33

Gambar 3.3 Komposisi tiap kategori pada tiap variabel independen 38

Gambar 3.4 Diagam pohon oleh metode CHAID pada kasus masa studi alumni 43

DAFTAR LAMPIRAN

Lampiran 1	Deskripsi Data	59
Lampiran 2	Output pemrosesan data dan pembentukan model dengan CHAID	62
Lampiran 3	Output pemrosesan data dan pembentukan model dengan regresi logistik biner	63
Lampiran 4	Output analisis regresi logistik biner dengan hanya memasukkan variabel signifikan terhadap model ……	67

BAB I

PENDAHULUAN

A. Latar Belakang

Klasifikasi merupakan salah satu bahasan yang sering kita dengar dalam statistika. Klasifikasi adalah proses untuk mengelompokkan suatu data menjadi kelompok-kelompok yang lebih kecil. Klasifikasi banyak digunakan dalam berbagai bidang, diantaranya bidang kesehatan dan pemasaran. Pada bidang kesehatan misalnya, klasifikasi bisa digunakan sebagai alat untuk mendiagnosa penyakit pasien dengan mengetahui ciri-ciri dari pasien tersebut. Sebagai contoh jika terdapat seorang pasien penderita kanker hati. Dengan melihat data dari pasien, dokter bisa memperkirakan stadium dari kanker yang diderita pasien tersebut. Data pasien yang dimaksud misalnya umur pasien, lama mengidap kanker, tingkat penyebaran sel kanker, jenis sel kanker, dll. Jika sudah diketahui berada pada stadium berapa kanker yang diidap oleh pasien, maka tim dokter bisa melakukan tindakan medis dan pengobatan yang lebih tepat.

Pada bidang pemasaran, klasifikasi bisa digunakan untuk mengetahui segmen-segmen/kelompok-kelompok dari konsumen. Jika telah diketahui segmen-segmen konsumen, maka perusahaan bisa memilih segmen mana yang tepat sebagai pangsa pasar dari produk yang mereka miliki. Misalnya ada sebuah bank ingin meningkatkan penyaluran produk pembiayaan mereka. Bank tersebut akan mempelajari data nasabah sebelumnya dari produk pembiayaan. Dengan membagi nasabah berdasarkan status kredit lancar atau macet, diketahui

kelompok-kelompok nasabah pembiayaan dengan latar belakang yang berbeda pada masing-masing kelompok. Pihak bank tinggal memutuskan kelompok nasabah mana yang dipilih sebagai sasaran utama pemasaran dengan mempertimbangkan status kredit. Diharapkan dengan penentuan kelompok sasaran pemasaran yang tepat, risiko kredit macet bisa diminimalkan.

Salah satu metode dalam tehnik klasifikasi yaitu metode CHAID (*Chi-Squared Automatic Interaction Detection)*. Metode CHAID merupakan metode yang relatif baru. Metode ini pertama kali diperkenalkan oleh oleh Dr. G. V. Kass pada tahun 1980 pada sebuah artikel berjudul *"An Exploratory Technique for Investigating Large Quantities of Categorial Data"* dalam buku *Applied Statistics.* Metode CHAID umumnya dikenal sebagai metode pohon klasifikasi (*Classification Tree Method*). Inti dari metode ini adalah membagi data menjadi kelompok-kelompok yang lebih kecil berdasarkan keterkaitan antara variabel dependen dengan variabel independen. Analisis CHAID digunakan ketika data yang dipakai adalah data dengan variabel-variabel kategorik. Variabel kategorik yaitu variabel yang memberikan label sesuai pengamatan dan dialokasikan untuk salah satu dari beberapa kemungkinan kategori, misalnya golongan darah O, A, B, AB (Everit & Skrondal, 2010)

Metode CHAID hanya efektif bila diterapkan pada data dengan pengamatan yang sangat banyak (Du Toit, S. H. C., A. G. W. Steyn & R. H. Stumph, 1986). Dibandingkan dengan metode klasifikasi untuk data dengan variabel kategorik lainnya seperti regresi logistik, CHAID memiliki kelebihan pada hasil *output*. Ketika regresi logistik hanya menampilkan persamaan regresi

yang digunakan untuk mengklasifikasikan data, CHAID akan menghasilkan output grafis berupa sebuah pohon klasifikasi sehingga membuat metode ini lebih mudah diintepretasikan karena bisa dilihat langsung bagaimana pola pemisahan dan penggabungan variabel independen pada prosesnya.

Sebagai contoh untuk melihat bagaimana penggunaan CHAID dalam klasifikasi, akan dipakai data alumni FMIPA UNY yang lulus antara bulan Desember 2004 sampai dengan Januari 2010. Variabel dependen yang digunakan adalah masa studi, dengan pengkategorian variabel tepat waktu dan tidak tepat waktu. Yang disebut tepat waktu adalah alumni yang menyelesaikan studinya kurang dari atau tepat 8 semester, dan yang lebih dari 8 semester disebut tidak tepat waktu. Metode CHAID nantinya akan membentuk segmentasi dari data alumni berdasarkan hubungan antara variabel dependen yaitu masa studi dengan variabel-variabel independen yang merupakan profil dari alumni. Hasil segmentasi akan menunjukkan segmen/kelompok alumni mana saja yang memiliki kelulusan tepat waktu ataupun tidak tepat waktu. Pihak kampus bisa menggunakan hasil segmentasi ini untuk mengetahui bagaimana ciri mahasiswa yang dikhawatirkan memiliki masa studi yang tidak tepat waktu dengan melihat ciri-ciri pada segmen alumni yang lulus tidak tepat waktu. Jika sudah diketahui tipe mahasiswa mana yang dikhawatirkan lulus tidak tepat waktu, pihak kampus bisa mengambil sebuah kebijakan yang tepat sehingga angka kelulusan tidak tepat waktu bisa dikurangi.

Sebagai analisis pembanding dari metode CHAID akan digunakan analasis yang sudah umum digunakan dalam klasifikasi data kategorik, yaitu analisis

regresi logistik. Regresi Logistik adalah bentuk khusus analisis regresi nonlinier yang memodelkan hubungan antara satu variabel dependen dengan sejumlah variabel independen. Berdasarkan variabel dependennya, regresi logistik dibagi menjadi regresi logistik biner dan regresi logistik ordinal. Disebut regresi logistik biner karena variabel dependen yang dipakai mempunyai dua nilai yang mungkin/ kategori, misalnya sukses/gagal, ya/tidak, lulus/tidak lulus. Sedangkan regresi logistik ordinal memiliki lebih dari dua kategori/ nilai yang mungkin pada variabel dependennya. Dalam kasus klasifikasi alumni FMIPA UNY, analisis yang dipakai adalah analisis regresi logistik biner karena variabel dependennya memiliki dua nilai yaitu lulus tepat waktu atau tidak tepat waktu.

B. Rumusan Masalah

Berdasarkan latar belakang di atas, permasalahan yang dirumuskan dalam penulisan skripsi ini adalah sebagai berikut :

1. Bagaimana prosedur analisis klasifikasi dengan metode CHAID?
2. Bagaimana hasil penerapan pada kasus klasifikasi alumni FMIPA UNY berdasarkan masa studi?
3. Bagaimana hasil perbandingan hasil klasifikasi metode CHAID dan regresi logistik biner pada kasus masa studi alumni FMIPA UNY?

C. Tujuan Penulisan

Dari rumusan masalah di atas, tujuan dari penulisan skripsi ini adalah :

1. Menjelaskan bagaimana prosedur metode CHAID sebagai salah satu metode dalam teknik klasifikasi.
2. Menerapkan metode CHAID untuk klasifikasi alumni FMIPA UNY berdasarkan masa studinya.
3. Membandingakan hasil klasifikasi metode CHAID dengan analisis regresi logistik biner.

D. Manfaat Penulisan

Manfaat yang diperoleh dari penulisan skripsi ini adalah :

1. Menambah wawasan tentang metode CHAID (*Chi-Squared Automatic Interaction Detection*) sebagai metode yang relatif baru dalam klasifikasi.
2. Informasi yang diperoleh dari kajian terapan dapat digunakan oleh fakultas untuk mengetahui bagaimana segmentasi mahasiswa FMIPA UNY berdasarkan masa studinya. Segmentasi mahasiswa adalah pembagian mahasiswa menjadi kelompok-kelompok yang lebih kecil yang memiliki ciri/karakteristik yang berbeda dan mungkin memerlukan perlakuan yang berbeda pula. Dengan diketahui segmen mahasiswa mana yang dikhawatirkan memiliki masa studi yang tidak tepat waktu, maka fakultas bisa merumuskan kebijakan yang tepat karena sudah mengetahui segmen sasaran agar jumlah mahasiswa tidak tepat waktu bisa dikurangi.

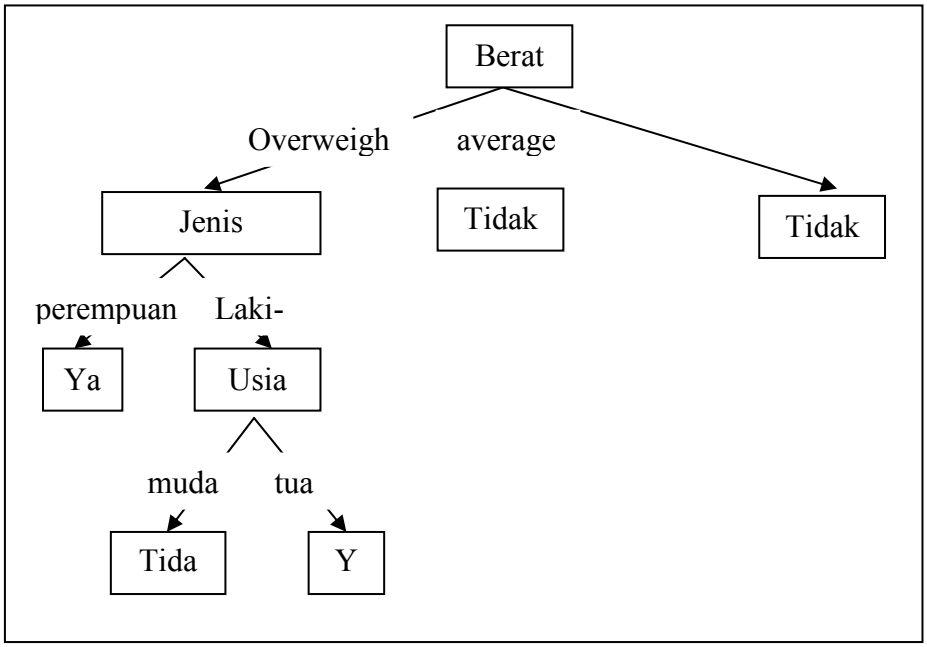
BAB II

KAJIAN PUSTAKA

A. Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang belum diketahui labelnya. Model ini sendiri bisa berupa aturan ”jika-maka”, berupa *decision tree*, formula matematis atau *neural network.*

Decision tree adalah salah satu metode klasifikasi yang paling popular karena mudah diintepretasi. Contoh *decision tree* dapat dilihat pada gambar berikut



Gambar 2.1 adalah contoh sebuah *decision tree* pada kasus klasifikasi penderita hipertensi. Pada gambar di atas setiap percabangan menyatakan kondisi yang harus dipenuhi dan tiap ujung pohon menyatakan kelas data. Dari gambar tersebut bisa diambil kesimpulan bahwa yang rentan terkena penyakit hiperetensi adalah orang yang memiliki kelebihan berat badan (*overweigt*) dan berjenis kelamin perempuan atau orang dengan kelebihan berat badan, laki-laki dan berusia lanjut/tua.

Proses klasifikasi biasanya dibagi menjadi dua fase : *learning* dan *test* (PBworks, 2007). Pada fase *learning*, sebagian data yang telah diketahui kelas datanya digunakan untuk membentuk model perkiraan Pada fase *test,* model yang sudah terbentuk diuji dengan sebagian data lainnya untuk mengetahui akurasi dari model tersebut. Bila akurasi mencukupi, model ini dapat dipakai untuk memprediksi kelas data yang belum diketahui.

B. Uji Independensi χ^2 (*Chi-Square*)

Uji χ^2 (*Chi-Square*) pertama dikembangkan oleh statistisi Inggris yang bernama Karl Pearson. Uji χ^2 (*Chi-Square*) antara lain dapat digunakan untuk mengetahui hubungan diantara dua variabel tertentu (untuk uji independensi), dimana variabel yang dimaksud mempunyai skala pengukuran nominal. Jika tidak terdapat hubungan antara variabel-variable tersebut, bisa dikatakan variabel-variabel tersebut bersifat independen atau saling bebas.

Misal suatu variabel pertama memiliki b kategori yaitu A_1, $A_{2,\ ...}$ A_b dan variabel kedua memiliki k kategori yaitu B_1, $B_{2,\ ...}$ $B_{k.}$ Banyak pengamatan pada ketegori ke-i (i = 1, 2, ..., b) variabel pertama dan kategori ke-j (j = 1, 2, ..., k) variabel kedua akan dinyatakan dengan O_{ij}. Hasilnya dapat dilihat dalam sebuah tabel kontingensi b x k sebagai berikut :

Tabel 2.1. Struktur data Uji *Chi Square*

Variabel 2 / **Variabel 1**	B_1	**…**	B_j	...	B_k	**Jumlah**
A_1	O_{11}		O_{1j}		O_{1k}	$n_{1\bullet}$
. . .						
A_i	O_{i1}		O_{ij}		O_{ik}	$n_{i\bullet}$
. . .						
A_b	O_{b1}		O_{bj}		O_{bk}	$n_{b\bullet}$
Jumlah	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet \mathbf{k}}$	n

Sumber **:** Haryatmi, S.(1986)

Keterangan :

O_{11} = Banyaknya pengamatan dengan sifat A_1 dan B_1

O_{ij} = Banyaknya pengamatan dengan sifat A_i dan B_j, i = 1, ... b, dan j = 1, ... k.

$n_{i\bullet}$ = Banyaknya pengamatan dengan sifat A_i, i = 1 ... b

$n_{\bullet j}$ = Banyaknya pengamatan dengan sifat B_j, j = 1, ... k

n $= \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$

Misalkan peluang kejadian A_i dan B_j adalah p_{ij} dengan i = 1,2, ... b, j = 1,2, ... k, maka peluang kejadian untuk untuk setiap A_i dan B_j dapat disajikan pada Tabel 2.2 berikut :

Tabel 2.2 Peluang kejadian

Kejadian 2 / **Kejadian 1**	B_1	...	B_j	**…**	B_k	**Jumlah**
A_1	p_{11}		p_{1j}		p_{1k}	$p_{1\bullet}$
⋮						
A_i	p_{i1}		p_{ij}		p_{ik}	$p_{i\bullet}$
⋮						
A_b	p_{b1}		p_{bj}		P_{bk}	$p_{b\bullet}$
Jumlah	$p_{\bullet 1}$		$p_{\bullet j}$		$p_{\bullet \mathbf{k}}$	

Sumber **:** Haryatmi, S.(1986)

Keterangan :

p_{ij} = Peluang kejadian A_i dan B_j

$p_{i\bullet}$ = Peluang total pada baris ke-i

$p_{\bullet \mathbf{j}}$ = Peluang total pada kolom ke-j

Nilai harapan untuk masing-masing sel adalah :

$$E_{ij} = n.p_{ij} = \frac{n_{i\bullet}\, n_{\bullet j}}{n} \qquad ; i = 1, ... b \text{ dan } j = 1, ... k \qquad ..... (2.1)$$

Statistik yang digunakan dalam alat uji hipotesis adalah :

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad ; i = 1, ... b \text{ dan } j = 1, ... k \qquad .....(2.2)$$

Statistik uji χ^2 berdistribusi *chi-square* dengan derajat bebas (b-1)(k-1). Berikut adalah langkah-langkah dalam uji hipotesis tersebut :

- Menuliskan hipotesis

  Ho : $p_{ij} = p_{i\bullet}\, p_{\bullet j}$ (kedua variabel independen)

  H_1 : $p_{ij} \neq p_{i\bullet}\, p_{\bullet j}$(kedua variabel tidak independen)

- Menentukan α
- Menentukan daerah penolakan, yaitu $\chi^2 > \chi^2_{\alpha;(b-1)(k-1)}$.
- Mencari $\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
- Mengambil kesimpulan
  a. Bila χ^2 masuk daerah penolakan, Ho ditolak
  b. Bila χ^2 tidak masuk dalam daerah penolakan, Ho diterima

C. Regresi Nonlinier

Analisis regresi ada dua macam, yaitu analisis regresi linier dan analisis regresi nonlinier. Nonlinier yang dimaksud pada regresi non linier bisa berarti nonlinier dalam variabel ataupun nonlinier dalam parameter. Fungsi regresi nonlinier dalam parameter adalah suatu fungsi apabila dideferensialkan hasilnya masih merupakan fungsi dalam parameter tersebut (Montgomery dan Peck, 1992). Model regresi logistik adalah contoh dari regresi nonlinier.

Suatu regresi bisa diketahui apakah memiliki kencederungan linier atau nonlinier dengan melihat diagram pencarnya. Apabila letak titik-titik objek dalam diagram XY (X sebagai variabel independen dan Y sebagai variabel dependen) berada di sekitar garis lurus, maka bisa diduga merupakan regresi linier. Jika letak titik-titik itu tidak di sekitar garis lurus, bisa lengkung, menyebar atau lainnya maka dapat diduga merupakan regresi nonlinier. Seperti pada regresi logistik, titik-titik objek hanya terpusat pada dua nilai Y=0 dan Y=1.

D. *Linier Probability Model* (LPM) dan Transformasi Logit

Penerapan regresi linier klasik pada variabel depeden kategorik merupakan asal mula lahirnya regresi logistik. Penerapan regresi linier klasik pada data dengan variabel dependen kategorik disebut sebagai *Linier probability Model*. Model regresi linier klasik ditulis dalam bentuk persamaan :

$$Y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \text{ ; E}(\varepsilon)=0 \qquad \ldots(2.3)$$

Variabel Y_i pada analisis regresi logistik merupakan variabel dikotom yang mempunyai dua nilai (0 atau 1) sedangkan Y_i pada regresi linier merupakan variabel kontinu, sehingga model regresi linier klasik (2.3) tidak bisa diterapkan secara langsung pada analisis regresi logistik.

Nilai ekspektasi dari Y_i yaitu :

$$E(Y_i) = E\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon\right)$$

$$E(Y_i) = \beta_0 + \beta_1 E(x_1) + \ldots + \beta_p E(x_p) + E(\varepsilon)$$

Karena $E(\varepsilon) = 0$ dan $E(x_1) = x_1$ maka $E(Y_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ …(2.4)

Karena Y_i pada regresi logistik biner hanya memiliki nilai 1 dan 0, maka nilai dari ekspektasi dari Y_i dapat dinyatakan sebagai :

$$E(Y_i=1) = \pi_i \qquad \ldots(2.5)$$

$$E(Y_i=0) = 1-\pi_i \qquad \ldots(2.6)$$

Karena Y_i diskrit, nilai ekspektasinya dapat dihitung dengan menggunakan rumus

$$E(Y_i) = \sum_{s=0}^{1} y_s f(y_s) = 1(\pi_i) + 0(1 - \pi_i) = \pi_{i,} \qquad \ldots(2.7)$$

Dari persamaan (2.4) dan (2.7) didapat *Linier Probability Model* sebagai berikut :

$$E(Y_i) = \pi_i = \beta_0 + \beta_i x_1 + \ldots + \beta_p x_p \qquad \ldots(2.8)$$

dengan :

$E(Y_i) = \pi_i$=nilai ekspektasi dari variabel Y_i

x_i = Variabel independen, i=1,2,…p

β_0= intersep regresi

β_i= koefisien regresi pada masing-masing x_i, i=1,2,…p

Dalam regresi logistik biner, nilai Y dikodekan dengan 1 dan 0. Kode 1 biasanya menyatakan status/peristiwa yang menjadi pokok perhatian (disebut *event*) misalnya kategori berhasil, sedangkan kode 0 menyatakan komplementnya (*non event*), yaitu gagal. Nilai Y bersifat diskrit, yaitu biner (hanya memiliki dua nilai), namun nilai π_i pada regresi logistik yang menyatakan nilai peluang dari Y_i bersifat kontinu karena nilainya dintentukan oleh kombinasi linier dari seluruh variabel independen X_i. Nilai π_i menyatakan peluang terjadinya *event* maupun *non event*, sehingga nilai π_i berkisar antara 0 dan 1.

Pada regresi linier, galat (□) diasumsikan berdistribusi normal sedangkan pada LPM galat □ diasumsikan berdistribusi binomial. Hal tersebut merupakan perbedaan mendasar antara model LPM dan model regresi linier klasik yang mengakibatkan beberapa permasalahan pada analisis LPM, diantaranya yaitu :

1. Galat □ berdistribusi binomial bukan berdistribusi normal, sehingga □ mempunyai dua nilai.

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$$

   Jika y_i =1 maka □$_i$ = $1 - (\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$
   Jika y_i = 0 maka □$_i$ = $-\beta_0 - \beta_1 x_1 - \ldots - \beta_p x_p$

2. Terdapat heteroskedastisitas yaitu variasi yang tidak konstan pada galat □. Galat model statistika seharusnya bersifat homoskedastik (variasi konstan). Misalkan didapat $\alpha^2(Y_i)$ dari model regresi linier klasik (2.3)

   $\alpha^2(Y_i) = E\{[Y_i - E(Y_i)]^2\} = (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i)$ atau

   $\alpha^2(Y_i) = \pi_i(1 - \pi_i) = [E(Y_i)][1 - E(Y_i)]$

   Variansi dari □$_i$ sama seperti pada y_i. Karena □$_i$=$y_i - \pi_i$ dan π_i adalah konstan, maka :

   $\alpha^2(\varepsilon_i) = \pi_i(1 - \pi_i) = [E(Y_i)][1 - E(Y_i)]$ atau

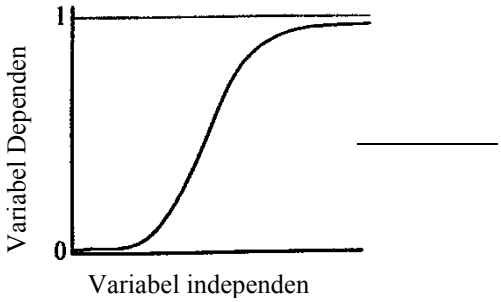
   $\alpha^2(\varepsilon_i) = (\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)(1 - \beta_0 - \beta_1 x_1 - \ldots - \beta_p x_p)$

   Terlihat bahwa $\alpha^2(\varepsilon_i)$ tergantung pada x_i. Oleh karena variansi dari galat error (□) tergantung pada x_i, metode *ordinary least square* tidak akan optimal. *Ordinary least square* yaitu metode untuk estimasi parameter pada analisis regresi dengan memperkecil perbedaan nilai respon yang diamati dengan nilai prediksi model.

3. π_i hanya menyatakan nilai peluang untuk variabel dependen yang dikotom dengan nilai antara 0 dan 1 yang mengakibatkan keterbatasan pada model peluang.

Untuk mengatasi permasalahan-permasalahan di atas dibutuhkan suatu transformasi yang memiliki sifat :

a. Seiring meningkatnya x, π juga meningkat tetap dalam interval [0,1]
b. Karena regresi logistik merupakan regresi nonlinier, maka relasi antara π dan x juga bersifat nonlinier



Gambar 2.2 Kurva regresi Logistik

Kurva berbentuk huruf S di atas hampir menyerupai CDF (Fungsi Distribusi Kumulatif) untuk sebarang variabel acak. Persamaan regresi logistik yang didasarkan pada CDF logistik yang dituliskan sebagai berikut :

…(2.9)

Hosmer dan Lemewshow(1989) menyebutkan bahwa untuk memudahkan dalam mengestimasi parameter β pada persamaan regresi logistik diperlukan transformasi logit , yang menghasilkan

…(2.10)

Pembuktian persamaan (2.10) adalah sebagai berikut :

Dari persamaan (2.9) nilai $\pi(x_i)$ adalah

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

Sehingga nilai $1-\pi(x_i)$ adalah sebagai berikut

$$1-\pi(x_i) = 1 - \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

$$\frac{\pi(x_i)}{1-\pi(x_i)} = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} \div \frac{1}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

$$\frac{\pi(x_i)}{1-\pi(x_i)} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

$$ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = ln\, e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

$$ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

E. Metode Maksimum *Likelihood*

Untuk mendapatkan persamaan regresi yang sesuai, nilai-nilai parameter yang terdapat pada model harus diestimasi terlebih dahulu. Metode estimasi yang digunakan dalam regresi logistik adalah Metode Maksimum *Likelihood*. Metode ini memiliki prinsip bahwa nilai estimasi yang digunakan adalah nilai estimasi yang memberikan nilai fungsi *Likelihood* yang paling besar.

Misalkan suatu sampel terdiri dari n observasi dari pasangan (X_1, Y_i), i=1,2,…,n. Dengan model regresi logistik $\pi(x_i) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$ setiap pasangan (X_1, Y_i) mempunyai fungsi kepadatan peluang :

$$f(x) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \qquad \text{…(2.10)}$$

Diasumsikan untuk setiap observasi, dapat dibentuk fungsi *Likelihood* sebagai berikut :

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

Dicari Ln *Likelihood*nya dengan persamaan :

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

$$= \sum_{i=1}^{n} \ln\,(\pi(x_i))^{y_i} + \sum_{i=1}^{n} \ln\,(1 - \pi(x_i))^{1-\mathrm{y_i}}$$

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

Dengan subsitusi $\pi(x_i) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$, dimana $g(x_i) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, maka diperoleh

$$L(\beta) = \sum_{i=1}^{n} y_i \ln\left[\frac{e^{g(x_i)}}{1 + e^{g(x_i)}}\right] + (1 - \mathrm{y_i})\ln\left[\frac{1}{1 + e^{g(x_i)}}\right]$$

$$= \sum_{i=1}^{n} \left[(y_i \ln\ e^{g(x_i)}) - (y_i \ln\ (1 + e^{g(x_i)})) - ((1 - y_i) \ln\ (1 + e^{g(x_i)})) \right]$$

$$= \sum_{i=1}^{n} (y_i \ln\ (\ e^{g(x_i)}) - \ln\ (1 + e^{g(x_i)}))$$

$$= \sum_{i=1}^{n} \left(y_i \left(g(x_i) \right) - \ln\ (1 + e^{g(x_i)}) \right)$$

$$= \sum_{i=1}^{n} \left(y_i \left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \right) - \ln\ (1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}) \right)$$

Untuk memperoleh nilai estimasi $\widehat{\beta_i}$, i=1,2,…,p yang memaksimumkan nilai fungsi $L(\beta)$, selanjutnya $L(\beta)$ dideferensialkan terhadap setiap β_i

$$L(\beta) = \sum_{i=1}^{n} \left(\beta_0 y_i + \beta_1 x_1 y_i + \cdots + \beta_p x_p y_i - \ln\ (1 + e^{\beta_0 y_i + \beta_1 x_1 y_i + \cdots + \beta_p x_p y_i}) \right)$$

$$\frac{\delta L(\beta)}{\delta \beta_0} = \sum_{i=1}^{n} \left[y_i - \left(\frac{1}{1 + e^{\beta_0 y_i + \beta_1 x_1 y_i + \cdots + \beta_p x_p y_i}} . e^{\beta_0 y_i + \beta_1 x_1 y_i + \cdots + \beta_p x_p y_i} \right) \right]$$

$$\frac{\delta L(\beta)}{\delta \beta_1} = \sum_{i=1}^{n} \left[x_1 y_i - \frac{1}{1 + e^{g(x_i)}} . x_1 e^{g(x_i)} \right]$$

dan

$$\frac{\delta L(\beta)}{\delta \beta_p} = \sum_{i=1}^{n} \left[x_p y_i - \frac{1}{1 + e^{g(x_i)}} . x_p e^{g(x_i)} \right]$$

Untuk mendapatkan β_i, i=1,2,…,p maka $\frac{\delta L(\beta)}{\delta \beta_p} = 0$

$$\frac{\delta L(\beta)}{\delta \beta_0} = \sum_{i=1}^{n} \left[y_i - \left(\frac{1}{1+e^{\beta_0 y_i + \beta_1 x_1 y_i + \cdots + \beta_p x_p y_i}} . e^{\beta_0 y_i + \beta_1 x_1 y_i + \cdots + \beta_p x_p y_i} \right) \right] = 0 \quad (2.11)$$

$$\frac{\delta L(\beta)}{\delta \beta_1} = \sum_{i=1}^{n} \left[x_1 y_i - \frac{1}{1+e^{g(x_i)}} . x_1 e^{g(x_i)} \right] = 0 \quad \ldots(2.12)$$

$$\frac{\delta L(\beta)}{\delta \beta_p} = \sum_{i=1}^{n} \left[x_p y_i - \frac{1}{1+e^{g(x_i)}} . x_p e^{g(x_i)} \right] = 0 \quad \ldots(2.13)$$

Persamaan (2.9),(2.10) dan (2.11) adalah persamaan yang tidak linier dalam β, sehingga untuk menyelesaikan persamaan-persamaan *Likelihood*nya sampai diperoleh estimasi dari β digunakan metode Newton-Raphson. Metode Newton-Raphson merupakan suatu metode untuk menyelesaikan persamaan nonlinier sehingga persamaan *Likelihood* pada regresi logistik dapat diselesaikan. Perhitungan dengan metode Newton-Raphson biasanya menggunakan bantuan komputer agar penyelesaiannya menjadi lebih cepat.

F. Metode Newton-Raphson

Estimasi parameter menggunakan metode maksimum *Likelihood* akan menghasilkan persamaan *Likelihood* yang nonlinier. Untuk menyelesaikan persamaan nonlinier tersebut agar didapatkan nilai estimasi parameter digunakan metode Newton-Raphson. Metode ini merupakan metode perhitungan iteratif sehingga akan lebih mudah jika dikerjakan dengan bantuan komputer.

Metode Newton_Raphson didasarkan pada deret Taylor (Pennington, 1967), sebagai berikut :

$$f(x_i+1) = f(x_i) + f'(x_i)(x_{i+1} - x_i) + \frac{f''(x_i)}{2!}(x_{i+1} - x_i)^2 + \cdots + \frac{f''(x_i)}{n!}(x_{i+1} - x_i)^n \quad \ldots(2.14)$$

Penyelesaian persamaan *Likelihood* dengan parameter θ menggunakan metode Newton-Raphson memperoleh nilai estimasi $\hat{\theta}$. Rumus estimasi parameter ke-(t+1) dalam proses iterasi (t=0,1,2,…) adalah sebagai berikut :

$$\hat{\theta}_{t+1} = \hat{\theta}_t - D(\hat{\theta}_t)^{-1} d(\hat{\theta}_t) \quad \ldots(2.15)$$

dengan $\hat{\theta}_{t+1}$ = estimasi parameter *θ* pada iterasi ke-(t+1)

$(\hat{\theta}_t)$ = estimasi parameter *θ* pada iterasi ke-t

d(θ) = matriks turunan pertama fungsi *Likelihood*, sehingga setiap entri dari *d(θ)* adalah $\frac{dL(\theta)}{d\theta}$

D(θ) = matriks turunan kedua fungsi *Likelihood*, sehingga setiap entri dari *D(θ)* adalah $\frac{d^2L(\theta)}{d\theta}$

Proses iterasi dengan menggunakan metode Newton-Raphson tidak berhenti sampai didapatkan nilai $\hat{\theta}$ yang konvergen yaitu sampai $\left|\frac{\hat{\theta}_{t+1} - \hat{\theta}_t}{\hat{\theta}_t}\right| < \delta$ dengan δ bilangan positif yang sangat kecil (Montgomery dan Peck,1992).

G. Inferensi dalam Regresi Logistik

Setelah estimasi dari *β* telah diketahui, langkah selanjutnya adalah inferensi parameter yang bisa dilakukan dengan uji Wald dan uji *Likelihood* rasio. Inferensi parmeter dilakukan untuk melihat apakah model regresi penuh (model

regresi setelah semua parameter dimasukkan ke dalam model) lebih baik daripada model regresi terreduksi (model regresi dengan hanya parameter β_0 yang dimasukkan ke dalam model). Model regresi logistik terreduksi adalah :

$$\pi(x_i) = \frac{e^{\beta_0}}{1+e^{\beta_0}} \qquad \ldots(2.16)$$

Pengujian keberartian model menggunakan Uji *Likelihood* rasio (*Likelihood Ratio Test*) yang merupakan pengujian terhadap parameter β_i secara simultan dengan hipotesis sebagai berikut :

H_0 : $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ (Model awal lebih baik dari model penuh)

H_1 : ada $\beta_i \neq 0$; i=1,2…,p (Model penuh lebih baik daripada model awal)

Statistik uji yang digunakan adalah statistik uji *G* di mana:

$$G = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_i}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\pi_i{}^{y_i}\,(1-\pi_i)^{1-y_i}}\right] \qquad \ldots(2.17)$$

Dengan $n_1 = \sum Y_i$, $n_0 = \sum(1-Y_i)$ dan $n = n_0 + n_1$

Jika H_0 benar, statistik uji *G* akan mengikuti sebaran distribusi χ^2 dengan derajat bebas (p-1) (Hosmer & Lemeshow, 1989). Hipotesis nol akan ditolak jika nila statistik uji $G > \chi^2(p-1,\alpha)$. Hipotesis nol ditolak artinya mode penuh (model dengan semua prediktor) lebih baik daripada model awal (model sebelum prediktor dimasukkan). Sedangkan pengujian parameter β_i secara parsial dilakukan dengan uji Wald dengan statistic uji W, yaitu :

$$W = \frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}}{S\hat{E}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}})} \qquad \ldots(2.18)$$

Dengan $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ sebagai penduga β_i dan SE ($\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$) sebagai penduga galat baku. Hipotesis yang akan diuji adalah :

H_o : β_i=0

H_1 : $\beta_i \neq 0$; i=0,1, …, p

Jika H_0 benar, maka statistic W akan mengikuti sebaran normal baku (Hosmer & Lemeshow, 1989). Keputusan tolak H_0 akan diambil jika $|W| > Z_{\alpha/2}$

H. Peluang Kejadian

Analisis regresi logistik biner mengklasifikasikan data ke dalam dua macam kategori yaitu *event* dan *non-event.*Kategori *event* yaitu kategori dengan nilai yang diharapkan misalnya kategori sukses, dan kategori *non-event* merupakan kategori dengan nilai yang tidak diharapkan misalnya kategori gagal. Misalkan *p* adalah peluang dikategorikannya sebuah data observasi ke dalam kategori *event*, maka peluang sebuah data observasi dikategorikan *non-event* adalah (p – 1).

I. Evaluasi Hasil Klasifikasi

Jika Analisis Regresi logistik (biner) digunakan untuk mengklasifikasikan data observasi, maka perlu diuji keakuratan fungsi regresi logistik dalam mengklasifikasikan observasi yang berasal dari sampel lain. Ketepatan hasil

klasifikasi dapat dihitung dari nilai *Apparent Error Rate* (APER) yang didefinisikan sebagai prosentase dari observasi yang salah dalam pengklasifikasian terhadap jumlah total observasi.

Tabel 2.3 *Confusius matrix*

		Predicted membership		
		Π_1	Π_2	
Actual membership	Π_1	n_{11}	n_{12}	n_1
	Π_2	n_{21}	n_{22}	n_2

Dengan Π_1 : kategori ke-1

Π_1 : kategori ke-2

n_1 : banyak observasi yang sesungguhnya termasuk dalam grup Π_1

n_2 : banyak observasi yang sesungguhnya termasuk dalam grup Π_2

n_{11} : banyak observasi yang termasuk dalam grup Π_1 dan diklassifikasikan dalam grup Π_1

n_{12} : banyak observasi yang termasuk dalam grup Π_1 dan diklassifikasikan dalam grup Π_2

n_{21} : banyak observasi yang termasuk dalam grup Π_2 dan diklassifikasikan dalam grup Π_1

n_{22} : banyak observasi yang termasuk dalam grup Π_2 dan diklassifikasikan dalam grup Π_2

APER dihitung dengan rumus :

$$APER = \frac{n_{12}+n_{21}}{n_1+n_2} \qquad \ldots(2.19)$$

Selain dengan menggunakan APER, keakuratan hasil klasifikasi juga dapat diketahui dengan menghitung Statistik uji Press's Q (Montgomery & Peck, 1992). Statistik uji ini bekerja dengan membandingkan jumlah observasi yang terklasifikasi dengan benar terhadap sampel total dan jumlah grup. Nilai Press's Q selanjutnya dibandingkan dengan nilai kritiknya, yaitu nilai *chi-square* dengan derajat bebas 1 pada tingkat signifikasi α. Ketepatan hasil klasifikasi dapat diterima jika Press's Q lebih besar dari nilai kritiknya.

$$Press's\ Q = \frac{[N-(nK)]^2}{N(K-1)} \qquad \ldots(2.20)$$

Dengan N : banyaknya keseluruhan sampel

n : Jumlah observasi yang terklasifikasi dengan benar

K : Banyak grup

BAB III

PEMBAHASAN

A. Analisis CHAID

Metode CHAID (*Chi-squared Automatic Interaction Detection*) pertama kali diperkenalkan pada sebuah artikel yang berjudul "*An Exploratory Technique for investigating Large Quantities of Categorical Data"* oleh Dr. G. V. Kass tahun 1980 pada buku *Applied Statistics.* Teknik tersebut merupakan teknik yang lebih awal dikenal sebagai *Automatic Interaction Detection (AID).* Metode CHAID secara umum bekerja dengan mempelajari hubungan antara variabel dependen dengan beberapa variabel independen kemudian mengklasifikasi sampel berdasarkan hubungan tersebut. Menurut Gallagher (2000), CHAID merupakan suatu teknik iteratif yang menguji satu-persatu variabel independen yang digunakan dalam klasifikasi, dan menyusunnya berdasarkan pada tingkat signifikansi statistik *chi-square* terhadap variabel dependennya.

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 variabel variabel independen yang lain. Proses tersebut terus berlanjut sampai tidak ditemukan lagi variabel-variabel independen yang signifikan secara statistik (Kunto dan Hasana, 2006)

CHAID adalah sebuah metode untuk mengklasifikasikan data kategori di mana tujuan dari prosedurnya adalah untuk membagi rangkaian data menjadi subgrup-subgrup berdasarkan pada variabel dependennya (Lehmann dan Eherler, 2001). Hasil dari pengklasifikasian dalam CHAID akan ditampilkan dalam sebuah diagram pohon.

CHAID tidak disarankan untuk data berukuran kecil. Penggunaan CHAID menjadi lebih berarti sejalan dengan meningkatnya banyak data yang dipakai. Du Toit, S. H. C., A. G. W. Steyn & R. H. Stumph (1986) menyebutkan bahwa banyak data minimal yang dipakai adalah 500.

Metode CHAID membagi data menjadi beberapa segmen berdasarkan hubungan variabel dependen dan variabel independennya. Variabel independen dalam metode CHAID akan dibedakan menjadi 3 bentuk yang berbeda yaitu monotonic, bebas dan mengambang (*float*).

1. Variabel-Variabel dalam Analisis CHAID

Variabel dependen dan independen dalam analisis CHAID adalah variabel kategorik. Menurut Gallagher (2000), CHAID akan membedakan variabel-variabel independen kategorik menjadi tiga bentuk yang berbeda, yaitu:

a. Monotonik

Yaitu variabel indenpenden yang kategori di dalamnya dapat dikombinasikan atau digabungkan oleh CHAID hanya jika keduanya

berdekatan satu sama lain atau mengikuti urutan aslinya (data ordinal). Contohnya: usia atau pendapatan.

b. Bebas

Yaitu variabel independen yang kategori di dalamnya dapat dikombinasikan atau digabungkan ketika keduanya berdekatan ataupun tidak(data nominal). Contohnya: pekerjaan, kelompok etnik, dan area geografis.

c. Mengambang (*floating*)

Yaitu variabel independen yang kategori di dalamnya dapat diperlakukan seperti monotonik kecuali untuk kategori yang *missing value*, yang dapat berkombinasi dengan kategori manapun.

2. Algoritma CHAID

Algoritma CHAID digunakan untuk melakukan pemisahan dan penggabungan kategori-kategori dalam variabel yang dipakai dalam analisisnya. Secara garis besar algoritma ini dapat dibagi menjadi tiga tahap, yaitu Penggabungan (*merging*), Pemisahan (*Splitting*) dan Penghentian (*Stopping*). Diagram pohon dimulai dari *root node* (node akar) melalui tiga tahap tersebut pada setiap *node* yang terbentuk dan secara berulang.

i) Penggabungan (*Merging)*

Tahap pertama dalam algoritma CHAID adalah penggabungan (*merging*). Pada tahap ini akan diperiksa signifikansi dari masing-masing kategori variabel independen terhadap variabel dependen. Tahap

penggabungan untuk setiap variabel independen dalam menggabungkan kategori-kategori non-signifikan adalah sebagai berikut :

1. Bentuk tabel kontingensi dua arah untuk masing-masing variabel independen dengan variabel dependennya.
2. Hitung statistik *chi-square* untuk setiap pasang kategori yang dapat dipilih untuk digabung menjadi satu, untuk menguji kebebasannya dalam sebuah sub tabel kontingensi 2 x J yang dibentuk oleh sepasang kategori tersebut dengan variabel dependennya yang mempunyai sebanyak J kategori. Langkah uji *chi-square* adalah sebagai berikut :
   - Menuliskan hipotesis

     Ho : $p_{ij} = p_{i\bullet}\, p_{\bullet j}$ (Variabel i dan variabel j independen atau tidak terdapat hubungan antara variabel i dan variabel j)

     H_1 : $p_{ij} \neq p_{i\bullet}\, p_{\bullet j}$ (Variabel i dan variabel j dependen atau terdapat hubungan antara variabel i dan variabel j)
   - Menentukan $\alpha = 0.05$

     Mencari harga $\chi^2_{0.05;(b-1)(k-1)}$ dari tabel *chi-square*.
   - Menentukan daerah penolakan, yaitu $\chi^2 > \chi^2_{0.05;(b-1)(k-1)}$ atau *p-value* < α
   - Mencari $\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
   - Mengambil kesimpulan

     a. Bila χ^2 masuk daerah penolakan, Ho ditolak

     b. Bila χ^2 tidak masuk dalam daerah penolakan, Ho diterima

Misalnya sebuah variabel independen X_i adalah variabel monotonik dengan a kategori, dimana i=1,2,...a. Variabel dependen Y memiliki b kategori. Untuk mengetahui kategori variabel independen mana yang tidak signifikan dipasangkan masing-masing kategori pada variabel independen dengan variabel dependen. Banyaknya pasangan yang mungkin adalah kombilasi b dari a.

Tabel 3.1 Ilustrasi pasangan penggabungan variabel

Kategori 1	Kategori 2	*p-value*
X_1	X_2	$p_{1,2}$
X_1	X_3	$p_{1,3}$
. .	. .	. .
X_a	X_1	$p_{a,1}$
. .	. .	. .
X_a	X_{a-1}	$P_{a,a-1}$

3. Untuk masing-masing nilai *chi-square* berpasangan, hitung *p-value* berpasangan bersamaan. Diantara pasangan-pasangan yang tidak signifikan, gabungkan sebuah pasangan kategori yang paling mirip (yaitu pasangan yang mempunyai nilai *chi-square* berpasangan terkecil dan *p-value* terbesar) menjadi sebuah kategori tunggal, dan kemudian dilanjutkan kelangkah nomor 4.

   Misalnya dari ilustrasi Table 3.1, pada tabel tersebut jika terdapat pasangan denga *p-value* lebih besar dari taraf signifikansinya, maka pasangan tersebut akan digabungkan. Misalnya pasangan kategori X_1 dan X_2 pada Tabel 3.1 tidak signifikan, maka pasangan tersebut akan digabungkan menjadi satu variabel baru yaitu $X_{1,2}$

4. Periksa kembali kesignifikansian kategori baru setelah digabung dengan kategori lainnya dalam variabel independen. Jika masih ada pasangan yang belum signifikan, ulangi langkah 3. Jika sudah semua sudah signifikan lanjutkan langkah berikutnya.

   Misalkan pada ilustrasi sebelumnya didapat gabungan variabel baru $X_{1,2}$. Variabel tersebut akan dipasangan dengan variabel lainnya misalnya X_3, X_4,...X_a kemudian dilihat apakah pasangan tersebut sudah signifikan, ketika semua signifikan bisa dilanjutkan ke langkah 5, namun jika masih ada yang belum signifikan kembali ke langkah 3.
5. Hitung *p-value* terkoreksi Bonferroni didasarkan pada tabel yang telah digabung.

ii) Pemisahan (*Splitting)*

Tahap *splitting* memilih variabel independen yang mana yang akan digunakan sebagai *split node* (pemisah node) yang terbaik. Pemilihan dikerjakan dengan membandingkan *p-value* (dari tahap merging) pada setiap variabel independen. Langkah *splitting* adalah sebagai berikut :

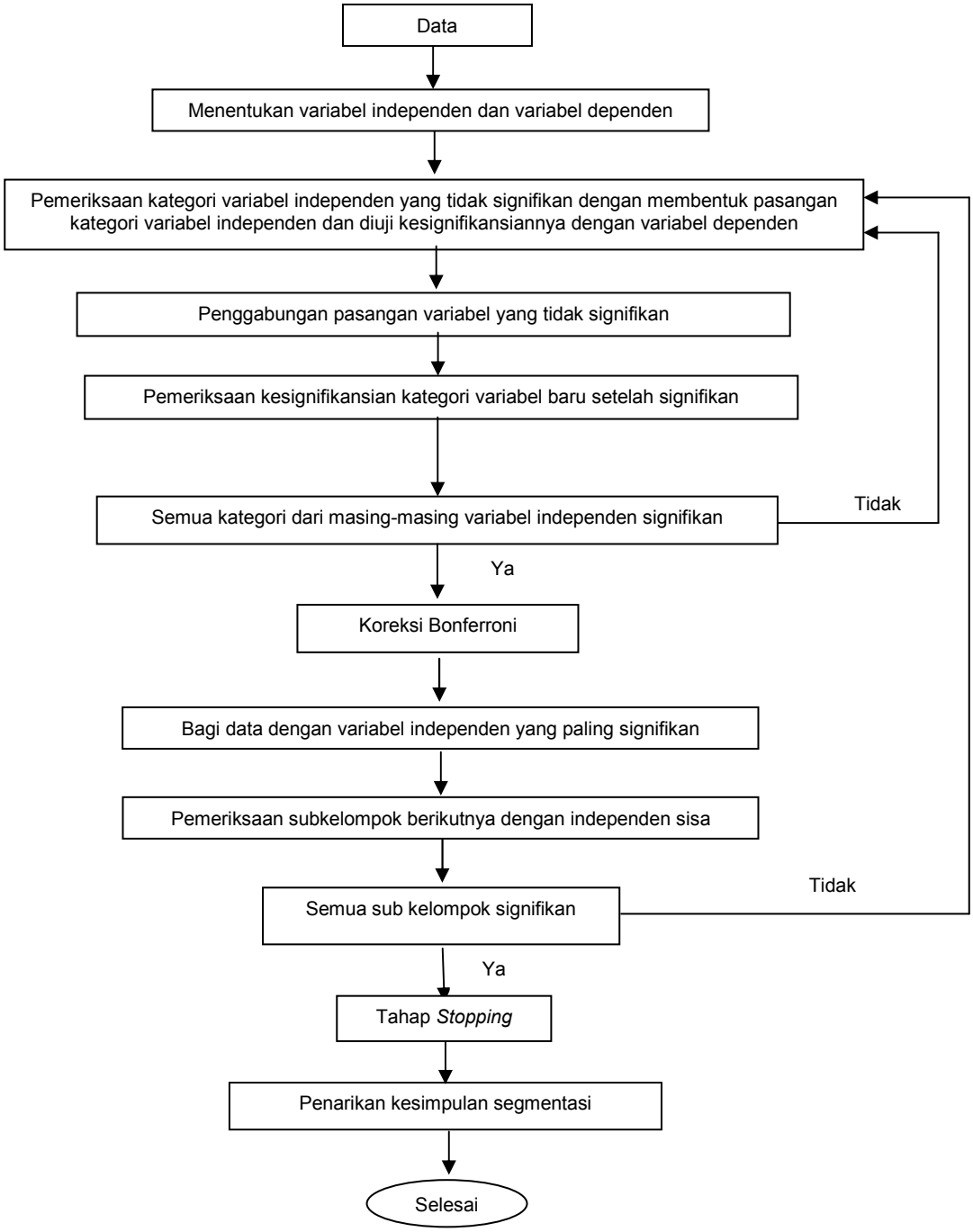
1. Pilih variabel independen yang memiliki *p-value* terkecil (paling signifikan) yang akan digunakan sebagai *split node*.
2. Jika *p-value* kurang dari sama dengan tingkat spesifikasi alpha, *split node* menggunakan variabel independen ini. Jika tidak ada variabel independen dengan nilai *p-value* yang signifikan, tidak dilakukan *split* dan *node* ditentukan sebagai *terminal node* (*node* akhir*)*

iii) Penghentian (*Stopping*)

Ulangi langkah penggabungan ntuk subkelompok berikutnya, Tahap *stopping* dilakukan jika proses pertumbuhan pohon harus dihentikan sesuai dengan peraturan pemberhentian di bawah ini :

1. Tidak ada lagi variabel independen yang signifikan menunjukkan perbedaan terhadap variabel dependen.
2. Jika pohon sekarang mencapai batas nilai maksimum pohon dari spesifikasi, maka proses pertumbuhan akan berhenti. Misalkan ditetapkan batas kedalaman pertumbuhan pohon klasifikasi adalah 3, ketika pertumbuhan pohon sudah mencapai kedalaman 3 maka pertumbuhan pohon klasifikasi dihentikan.
3. Jika ukuran dari *child node* kurang dari nilai ukuran *child node* minimum spesifikasi, atau berisi pengamatan-pengamatan dengan banyak yang terlalu sedikit maka *node* tidak akan di-*split*. Misalkan ditetapkan ukuran minimal *child node* adalah 50, ketika *splitting* mengasilkan ukuran *child node* kurang dari 50, maka node tersebut tidak akan dipecah.

Langkah-langkah pada algoritma CHAID dapat digambarkan spada diagram alir pada Gambar 3.1.



Gambar 3.1. Diagram Alir Algoritma CHAID

3. Koreksi Bonferroni (*Bonferroni Correction*)

Andaikan bahwa variabel independen memiliki c kategori dan

dikurangi menjadi r kategori pada langkah penggabungan, maka perkalian *Bonferroni* adalah banyaknya cara yang mungkin yang mana c kategori dapat digabungkan menjadi r kategori. Dengan demikian nilai *p-value* dari uji *chi-square* untuk independensi yang baru merupakan perkaliannya dengan pengali *bonferroni* sesuai dengan jenis variabelnya (Gallagher, 2000).

Koreksi Bonferroni adalah suatu proses koreksi yang digunakan ketika beberapa uji statistik untuk kebebasan atau ketidakbebasan dilakukan secara bersamaan (Kunto dan Hasana, 2006). Koreksi Bonferroni biasanya digunakan dalam perbandingan berganda.

Gallagher (2000) menyebutkan bahwa pengali *Bonferroni* untuk masing-masing jenis variabel independen adalah sebagai berikut:

1. Variabel independen Monotonik

$$M = \binom{c-1}{r-1} \qquad \text{..... (3.1)}$$

dimana :

M = Pengali Bonferroni

c = banyaknya kategori variabel independen awal

r = banyaknya kategori variabel independen setelah penggabungan

2. Variabel independen Bebas

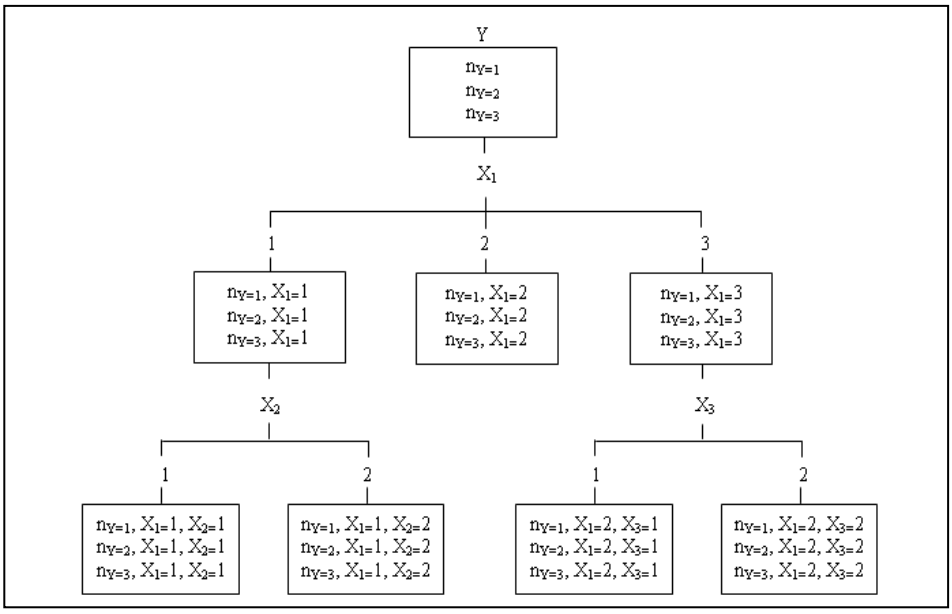
$$M = \sum_{i=0}^{r-1} (-1)^i \frac{(r-1)^c}{i!(r-i)!} \qquad \text{..... (3.2)}$$

3. Variabel independen Mengambang *(Floating*)

$$M = \binom{c-2}{r-2} + r\binom{c-2}{r-1} \qquad \text{..... (3.3)}$$

4. Diagram Pohon Klasifikasi CHAID (CHAID *Classification Tree*)

CHAID akan menghasilkan sebuah diagram pohon klasifikasi yang menggambarkan pembentukan segmen. Diagram pohon CHAID ditunjukkan pada gambar 3.1. Diagram CHAID terdiri dari batang pohon (*tree trunk)* dengan membagi (*split)* menjadi lebih kecil berupa cabang-cabang (*brances).*



Gambar 3.2. Diagram Pohon dalam Analisis CHAID
Sumber : *Responder Profiling with CHAID and Dependency Analysis*, Lehmann, T. dan Eherler, D. 2001.

Menurut Myers (Kunto dan Hasana, 2006), diagram pohon CHAID mengikuti aturan "dari atas ke bawah" (*Top-down stopping rule*), dimana diagram pohon disusun mulai dari kelompok induk (*parent node)*, berlanjut di bawahnya sub kelompok (*child node)* yang berturut-turut dari hasil pembagian kelompok induk berdasarkan kriteria tertentu. Node pada ujung pohon yang tidak terdapat percabangan lagi disebut *terminal node*. Tiap-tiap *node* dari diagram pohon ini menggambarkan

sub kelompok dari sampel yang diteliti dan berisi keseluruhan sampel dan frekuensi absolut n_i untuk setiap kategori yang disusun.

Pada pohon klasifikasi CHAID terdapat istilah kedalaman (*depth*) yang berarti banyaknya tingkatan *node-node* sub kelompok sampai ke bawah pada *node* sub kelompok yang terakhir. Pada kedalaman pertama, sampel dibagi oleh X_1 sebagai variabel independen terbaik untuk variabel dependen berdasarkan uji *chi-square*. Tiap *node* berisi informasi tentang frekuensi variabel Y, sebagai variabel dependen, yang merupakan bagian dari sub kelompok yang dihasilkan berdasarkan kategori yang disebutkan (X_1). Pada kedalaman ke-2 (*node* X_2 dan X_3) merupakan pembagian dari X1 (untuk *node* ke-1 dan ke-3). Dengan cara yang sama, sampel selanjutnya dibagi oleh variabel penjelas yang lain, yaitu X_2 dan X_3 , dan selanjutnya menjadi sub kelompok pada *node* ke-4, 5, 6, dan 7 (Lehmann dan Eherler, 2001). Pada masing-masing *node* ditampilkan persentase responden untuk setiap kategori dari variabel dependen, dan juga ditunjukkan jumlah total responden untuk masing-masing *node*.

B. Klasifikasi Alumni FMIPA UNY

Klasifikasi alumni S1 FMIPA UNY akan digunakan sebagai contoh penerapan CHAID dalam klasifikasi. Data alumni diambil dari Subbag Sistem Informasi Universitas Negeri Yogyakarta. Sedangkan data yang dipakai adalah data alumni S1 FMIPA UNY yang lulus antara bulan Desember 2004 sampai dengan bulan januari 2010. Dalam data tersebut terdapat dua macam variabel,

yaitu variabel dependen dan independen. Varibel dependen yang digunakan adalah masa studi, sedangkan variabel independennya yaitu jenis kelamin, asal daerah, jalur masuk, program studi dan IP semester pertama. Berikut adalah definisi variabel-variabel tersebut :

i) Masa studi (Y)

Masa Studi adalah banyak semester yang ditempuh alumni pada saat masih menempuh pendidikan di FMIPA UNY. Masa studi ideal bagi mahasiswa S1 adalah 8 semester. Jika menempuh lebih dari 8 semester berarti bisa dikatakan melebihi masa studi ideal atau tidak tepat waktu. Skala data untuk variabel masa studi termasuk dalam skala nominal yang dibedakan menjadi dua kategori, yaitu tepat waktu dengan kode 1 dan tidak tepat waktu dengan kode 2.

ii) Jenis Kelamin

Skala data untuk variabel jenis kelamin termasuk dalam skala nominal. Jenis Kelamin alumni dikategikan laki-laki dan perempuan, masing-masing diberi kode secara berurutan 1 dan 2.

iii) Asal Daerah

Asal daerah dikelompokkan menjadi 3 kategori, karena mahasiswa FMIPA berbagai kota di Indonesia, sehingga tidak mungkin disebutkan satu persatu tanpa pengkategorian, berikut pengkategorinya :

Tabel 3.2 Pengkategorian alumni dan pemberian kode berdasarkan asal daerahnya

Asal Daerah	Kode
DIY	1
Pulau Jawa (selain DIY)	2
Luar Pulau Jawa	3

Skala data untuk variabel asal daerah termasuk skala data nominal, sehingga variabel independen asal daerah disebut variabel independen bebas.

iv) Jalur Masuk

Jalur masuk S1 FMIPA UNY dibedakan menjadi dua kategori, yaitu jalur Reguler bagi mahasiswa yang masuk melalui SNMPTN dan PBU dengan kode 1, dan jalur non regular bagi mahasiswa yang masuk melalui selain dengan jalur reguler dengan kode 2. Skala data untuk variabel jalur masuk termasuk skala data nominal, sehingga variabel independen jalur masuk disebut variabel independen bebas.

v) Program studi

UNY yang dahulu disebut IKIP sebagai universitas kependidikan memiliki jurusan-jurusan kependidikan. Namun semenjak perubahan dari institut menjadi universitas, UNY juga memiliki prodi non-kependidikan atau biasa disebut murni. Di FMIPA saat ini terdapat 5 jurusan, yaitu pendidikan matematika, pendidikan kimia, pendidikan fisika, pendidikan biologi dan pendidikan IPA. Pada 4 jurusan yang disebutkan pertama, masing-masing jurusan memiliki prodi kependidikan dan non-kependidikan. Data alumni yang dipakai dalam analisis kali ini adalah data alumni yang pengkategorian program studinya berdasarkan berasal dari program studi pendidikan atau non-kependidikan. Pengkodean prodi pendidikan adalah 1 dan non kependidikan adalah 2. Skala data untuk variabel program studi termasuk skala data nominal,

sehingga variabel independen program studi disebut variabel independen bebas.

vi) IPK semester 1

Indeks Prestasi semester 1 akan dijadikan sebagai salah satu veriabel independen dengan pembagian menjadi 4 kategori, yaitu :

Tabel 3.3 Pengaktegorian alumni dan pemberian koden berdasarkan IPK semester 1

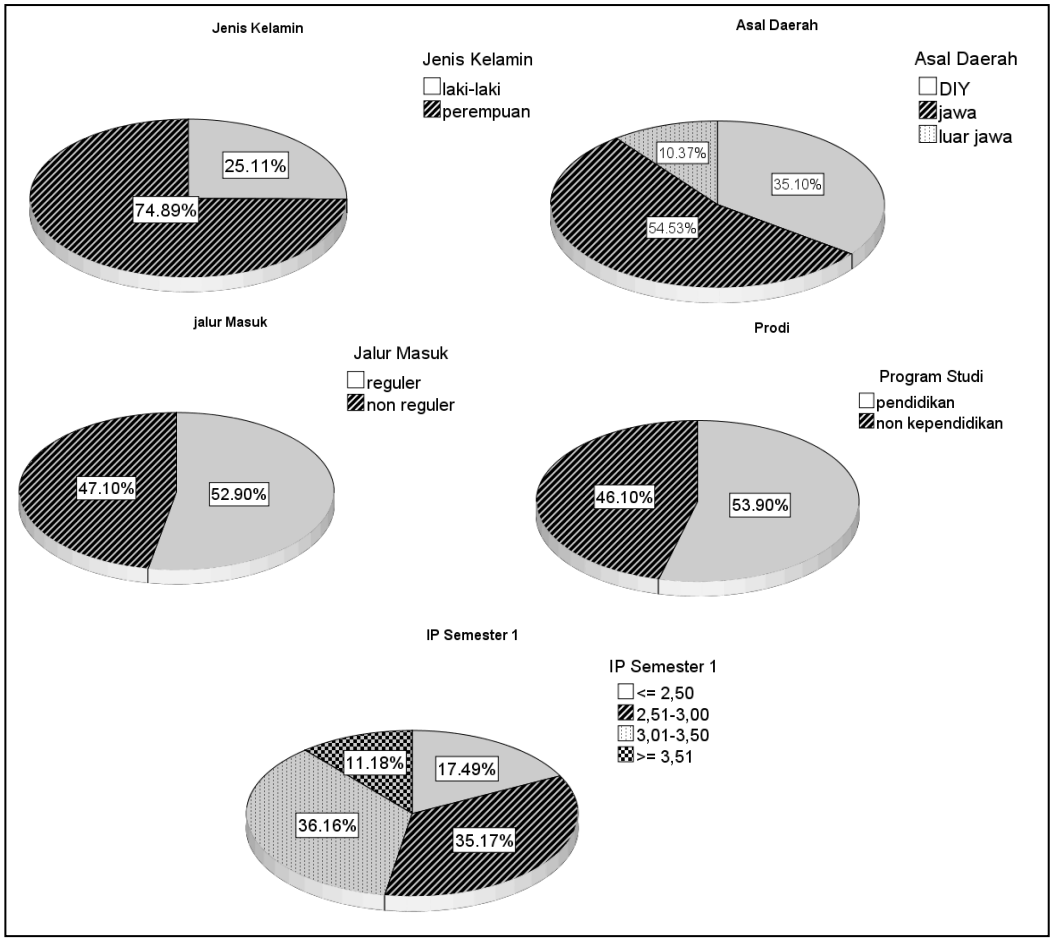
IP sem 1	Kode
≤ 2,50	1
2,51 – 3,00	2
3,01 – 3,50	3
≥ 3,51	4

Skala data untuk variabel IPK semester 1 termasuk dalam skala ordinal, sehingga variabel independen IPK semester 1 disebut variabel independen monotoik.

1. Deskripsi Variabel

Data dikumpulkan dari 1601 alumni, dan digambarkan sebagai *pie chart* dari variabel-variabel independennya pada gambar 3.2. Berdasarkan gambar 3.2 dapat dilihat bahwa alumni FMIPA UNY antara desember 2004 sampai dengan januari 2010 sebagian besar adalah perempuan. Asal daerah alumni FMIPA sebagian besar berasal dari luar DIY namun masih di pulau. Berdasarkan jalur masuk, antara jalur regular dan non regular hampir berimbang jumlahnya, yaitu 847 untuk regular dan 754 dari non regular. Alumni berasal dari 2 macam program studi, yaitu pendidikan dan non kependidikan, jumlah dari masing-masing kategori prodi hampir berimbang. Kemudian berdasarkan IP semester pertama, sebagian besar memiliki IP 3,01-3,50 pada semester 1, Untuk perincian

jumlah dari tiap kategori dari masing-masing variabel dapat dilihat pada lampiran 1.



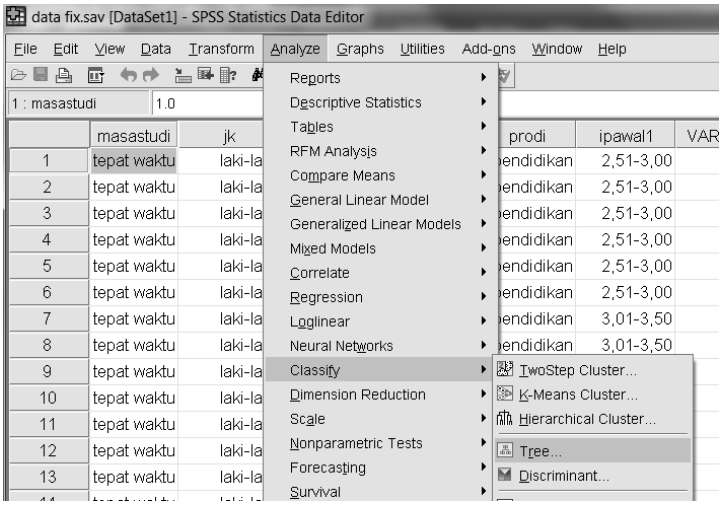
Gambar 3.3. Komposisi tiap kategori pada setiap variabel independen

2. Analisis Data dengan metode CHAID

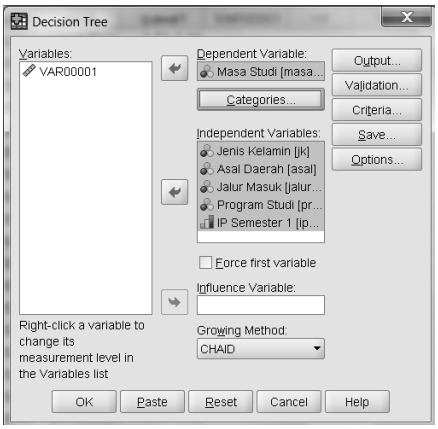
i) Metode Analisis Data

Pengolahan data dan pembuatan pohon klasifikasi untuk kasus klasifikasi alumni FMIPA UNY berdasarkan masa studinya menggunakan alat bantu program komputer *Statistical Product and Service Solution* (SPSS) *version* 17. Berikut adalah langkah-langkah penggunaan program SPSS untuk klasifikasi metode CHAID :

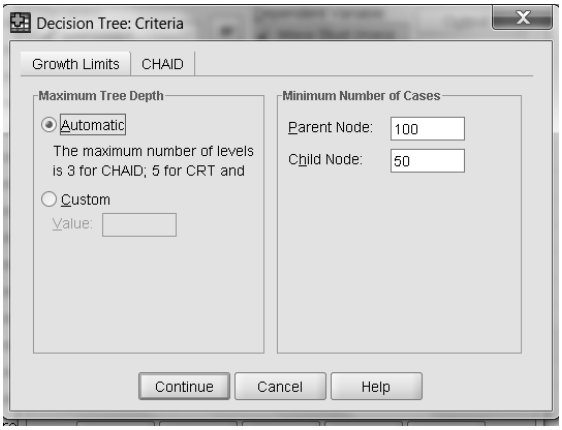
1) Pada *toolbar* program SPSS klik *Analyse* → *Classify* → *Tree*



2) Pada jendela *Decision Tree*, masukkan variabel dependen dan independen kemudian pada *Growing Method* pilih CHAID



3) Pada Jendela *Decision Tree* klik *button Criteria* dan akan muncul jendela *Criteria*. Masukkan jumlah data minimum yang dikehendaki pada *parent node* dan *child node* kemudian klik *continue*.



4) Setelah kembali ke jendela *Decision Tree* klik tombol *OK* dan kemudian akan muncul output metode CHAID.

ii) Intepertasi output SPSS

Tabel *model summary* (ringkasan model) pada Tabel 3.3 menyediakan beberapa informasi tentang spesifikasi metode CHAID yang digunakan. Tabel 3.3 bagian *specification* menunjukkan bagaimana pengaturan yang dipakai untuk membangun model pohon klasifikasi, juga termasuk di dalamnya variabel-variabel yang digunakan dalam analisis.

Pada variabel independen, prosedur secara otomatis mengeluarkan variabel independen yang tidak secara signifikan memberikan kontribusi/pengaruh pada model akhir. Oleh karena itu tarlihat di tabel, pada *specification* terdapat 5 variabel independen, sedangkan pada bagian *result* hanya terdapat 4 variabel independen saja yang secara signifikan memberikan kontribusi kepada model akhir. Variabel independen tersebut adalah IP semester 1, Program Studi, Jalur Masuk, dan Asal Daerah. Sedangkan variabel Jenis Kelamin tidak dimasukkan karena tidak signifikan berkontribusi pada model akhir.

Tabel 3.4 Ringkasan pembentukan model degan metode CHAID

Model Summary

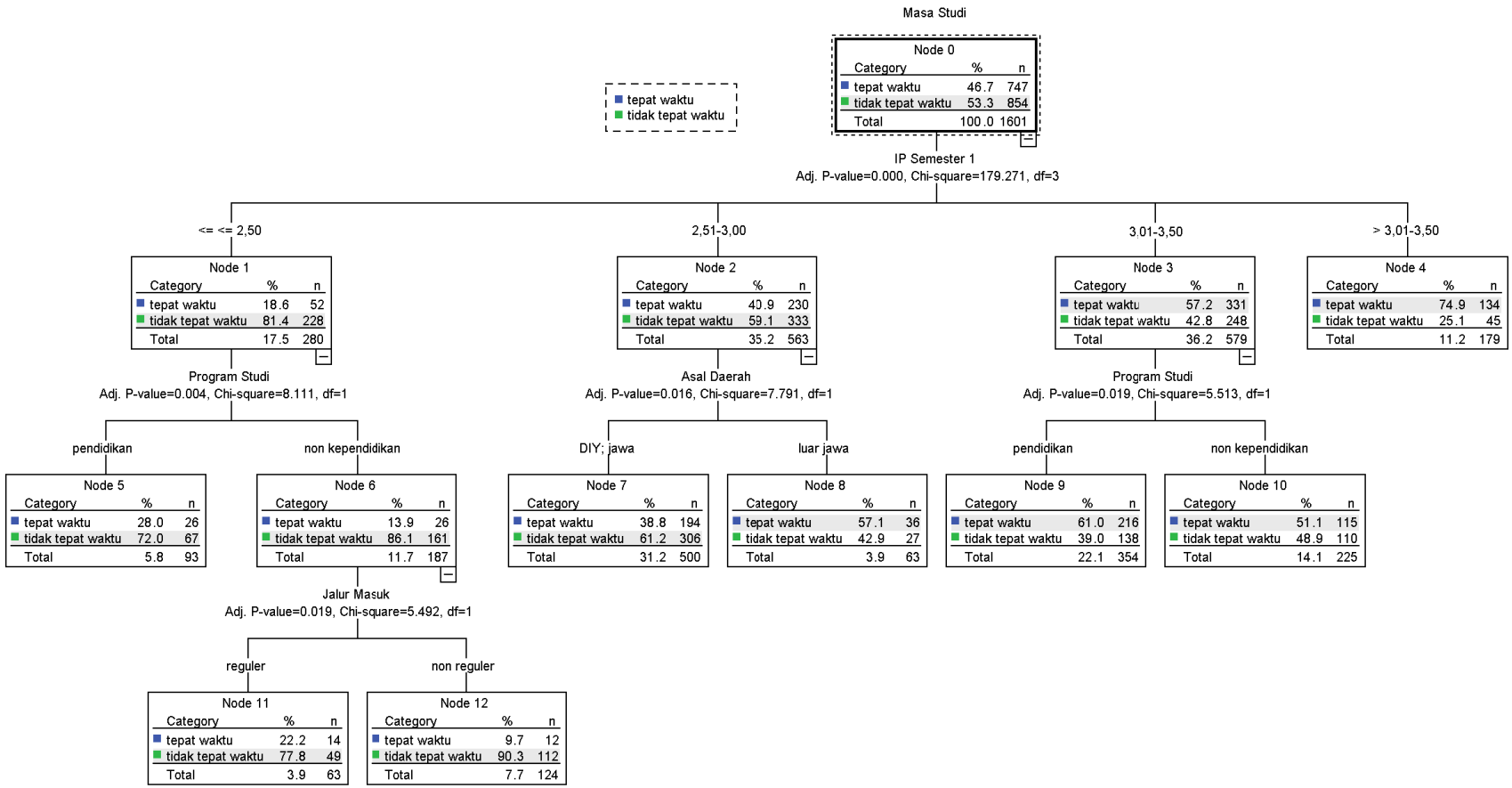
Specifications	Growing Method	CHAID
	Dependent Variable	Masa Studi
	Independent Variables	IP Semester 1, Jenis Kelamin, Asal Daerah, Jalur Masuk, Program Studi
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	IP Semester 1, Program Studi, Jalur Masuk, Asal Daerah
	Number of Nodes	13
	Number of Terminal Nodes	8
	Depth	3

Maximum tree depth pada bagian *specification* adalah untuk membatasi pertumbuhan pohon klasifikasi. Pada tabel *model summary* tersebut nilai *maximum tree depth* adalah 3 yaitu nilai batas yang sudah ditetapkan secara otomatis oleh program SPSS. Sedangkan *minimum cases in parent node* (jumlah minimum kasus pada *parent node*) sebanyak 100 dan *minimum cases in child node* (jumlah minimum kasus pada *child node*) sebanyak 50. Keduanya juga merupakan nilai otomatis yang diberikan oleh SPSS.

Hasil pembentukan model berupa pohon klasifikasi dengan node sebanyak 13 buah, node akhir (*terminal node*) sebanyak 8 buah dengan level kedalaman (*depth*) sebanyak 3. Variabel IP semester 1 merupakan variabel independen yang

paling signifikan terhadap model, hal itu bisa dilihat pada gambar 3.4 dimana variabel IP semester 1 berada pada *node-node* awal, yaitu pada *node* 2 sampai 5. Dari diagram pohon klasifikasi pada Gambar 3.3 dapat diintepretasikan sebagai berikut :

i. Sesuai dengan tabel model summary, diagram pohon tersebut memiliki 13 node, dengan jumlah terminal node sebanyak 8.
ii. Kedalaman dari pohon klasifikasi tersebut adalah 3.
iii. Diketahui variabel yang signifikan dalam membentuk model adalah IP semester 1, Asal Daerah, Jalur masuk, Program studi.
iv. Masing-masing *terminal node* dapat diartikan sebagai salah satu segmen dari alumni FMIPA UNY, segmen-segmen tersebut dapat dilihat pada Tabel 3.4.
v. Penarikan kesimpulan segmen alumni yang paling besar kelulusan tidak tepat waktu dan kelulusan tepat waktunya bisa dilihat dari persentase kategori lulus tepat waktu dan lulus tidak tepat waktu pada masing-masing *node*.



Gambar 3.4 Diagram pohon oleh metode CHAID pada kasus masa studi alumni

Tabel 3.5 Segmentasi Alumni FMIPA UNY

Segmen ke-1	Alumni yang memiliki IP semester 1 kurang dari sama dengan 2,50 dan program studi pendidikan
Segmen ke-2	Alumni yang memiliki IP semester 1 kurang dari sama dengan 2,50, program studi non kependidikan dan jalur masuk non regular
Segmen ke-3	Alumni yang memiliki IP semester kurang dari sama dengan 2,50, program studi non kependidikan dan jalur masuk reguler
Segmen ke-4	Alumni yang memiliki IP semester 1 antara 2,51-3,00, serta asal daerah DIY dan dari luar DIY tapi masih dari pulau jawa
Segmen ke-5	Alumni yang memiliki IP semester 1 antara 2,51-3,00, dan berasal dari luar jawa
Segmen ke-6	Alumni yang memiliki IP semester 1 antara 3,01-3,50 serta berasal dari prodi pendidikan
Segmen ke-7	Alumni yang memiliki IP semester 1 antara 3,01-3,50 serta berasal dari prodi prodi non kependidikan
Segmen ke-8	Alumni yang memiliki IP semester 1 >3,01-3,50, atau dengan kata lain lebih besar atau sama dengan 3,51

Tabel 3.6 Persentase setiap segmen alumni FMIPA UNY

Segmen	Masa studi tepat waktu		Masa studi tidak tepat waktu	
	Jumlah alumni	Persentase	Jumlah alumni	Persentase
1	26	28 %	67	72 %
2	12	9,7 %	112	90,3 %
3	14	22,2 %	49	77,8 %
4	194	38,8 %	306	61,2 %
5	36	57,1 %	27	42,9 %
6	216	51 %	138	39 %
7	115	51,1 %	110	48,9 %
8	134	74,9 %	45	25,1 %

Dari Tabel 3.5 bisa dilihat bahwa persentase terbesar alumni yang memiliki masa studi tepat waktu adalah pada kelompok/segmen ke-8, yaitu alumni yang memiliki IP semester 1 lebih dari atau sama dengan 3,51 yaitu dengan persentase 74,9%. Sedangkan kelompok/segmen yang memiliki masa studi tidak tepat waktu dengan prosentase 90,3% adalah segmen ke-2, yaitu alumni yang IP semester 1 kurang dari sama dengan 2,50, program studi non kependidikan dan berasal dari jalur masuk non regular.

Dari hasil metode CHAID tersebut telah diketahui kelompok-kelompok/segmen-segmen yang terbentuk dari data alumni FMIPA UNY. Hasil dari klasifikasi ini bisa juga digunakan untuk mengetahui kelompok-kelompok/segmen-segmen pada mahasiswa yang masih aktif. Sudah diketahui dari data alumni bahwa kelompok/segmen ke-2 memiliki masa studi yang tidak tepat waktu, maka pada mahasiswa yang masih aktif yang memiliki ciri-ciri seperti pada segmen tersebut perlu diwaspadai kalau mereka memiliki kecenderungan untuk terlambat dalam studinya/ memiliki masa studi yang tidak tepat waktu.

Tabel 3.7 Klasifikasi dari metode CHAID

Observasi	Prediksi		
	tepat waktu	tidak tepat waktu	Persentase benar
tepat waktu	501	246	67.1%
tidak tepat waktu	320	534	62.5%
Persentase total	51.3%	48.7%	64.6%

Tabel risiko (pada lampiran 2) dan klasifikasi menyebutkan evaluasi dari seberapa bagus model bekerja. Estimasi risiko sebesa 0,354 menunjukkan bahwa prediksi kategori oleh model (masa studi alumni, tepat waktu atau tidak tepat waktu) jelek untuk 35,4% kasus. Sehingga risiko dari klasifikasi yang salah untuk masa studi alumni maksimal sebesar 35,4%.

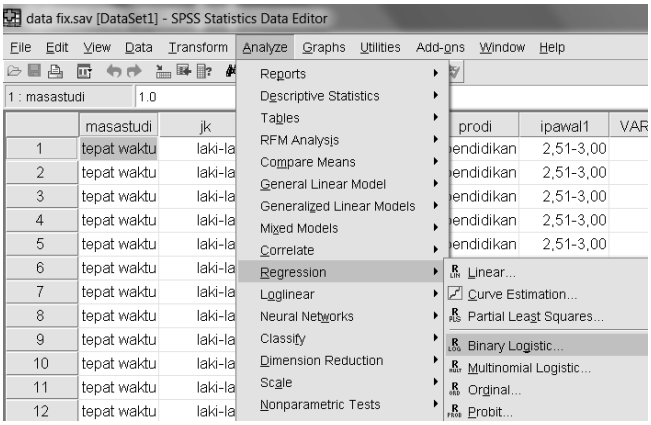
Sesuai dengan tabel risiko, tabel klasifikasi juga menunjukkan nilai yang senada. Prersentase model mengklasifikasikan masa studi alumni secara tepat sebesar 64,6 %

3. Analisis Data dengan Regresi Logistik Biner

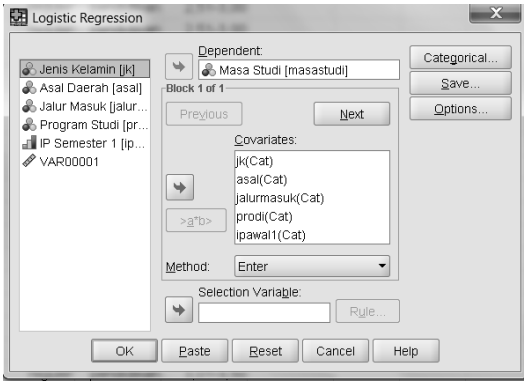
i) Metode Analisis Data

Pengolahan data dalam analisis regresi logistik biner kembali akan menggunakan bantuan program komputer *Statistical Product and Service Solution* (SPSS) *version* 17 dengan langkah-langkah sebagai berikut :

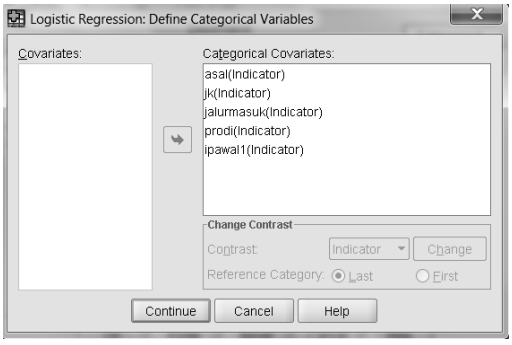
1) Pada *toolbar* program SPSS klik A*nalyse* → *Regression* → *Binary Logistic*



2) Pada jendela *Logistic Regression* masukkan variabel masa studi pada *Dependent* dan variabel yang lain pada *covariates*



3) Masih pada jendela *Logistic Regression*, klik *Categorical* dan akan muncul jendela *Define Categorical Variable*. Masukkan semua variabel pada kotak sebelah kiri pada kotak *Categorical Covariates* kemudian klik *continue.*



4) Setelah kembali pada jendela *Logistic Regression*, klik tombol *OK* dan kemudian program SPSS akan menampilkan *output* dari analisis regresi logistik biner.

ii) Pembentukan Variabel Semu

Pada kasus ini, semua variabel independen merupakan data kategori, sehingga pemilihan model dengan variabel semu melibatkan semua variabel independennya. Berikut ini adalah tabel pembentukan variabel semu :

Tabel 3.8 Pembentukan Variabel Semu

			Parameter coding		
		Frequency	(1)	(2)	(3)
IP Semester 1	<= 2,50	280	.000	.000	.000
	2,51-3,00	563	1.000	.000	.000
	3,01-3,50	579	.000	1.000	.000
	>= 3,51	179	.000	.000	1.000
Asal Daerah	DIY	562	1.000	.000	
	Jawa	873	.000	1.000	
	luar jawa	166	.000	.000	
Jalur Masuk	Regular	847	1.000		
	non reguler	754	.000		
Jenis Kelamin	laki-laki	402	1.000		
	perempuan	1199	.000		
Program Studi	pendidikan	863	1.000		
	non kependidikan	738	.000		

iii) Pembentukan Persamaan Regresi Logistik

Tabel 3.9 Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.134	.050	7.140	1	.008	1.143

Dari tabel di atas, persamaan regresi logit awal adalah $g(x) = \beta_0 = 0{,}134$

Persamaan di atas hanya memuat konstanta saja. Untuk mendapatkan nilai β_i digunakan metode maksimum *Likelihood* yang dapat kita lihat hasilnya dari *output* program SPSS tabel *Iteration History* pada lampiran 3.

Dengan melihat nilai signifikansi (Sig) pada tabel Variables in the Equation(2) yang terlampir pada lampiran 3, variabel yang signifikan dalam persamaan regresi logistik adalah variabel ipawal1(1), ipawal1(2), ipawal1(3). Model logitnya dapat dinyatakan sebagai berikut :

g(x) = 1,045 – 1,054[ipawal(1)] – 1,665[ipawal1(2)] - 2,450[ipawal1(3)] - 0,447[prodi(1)]

dengan memisalkan variabel ipawal(1) sebagai X_1, ipawal(2) sebagai X_2, ipawal(3) sebagai X_3 dan prodi(1) sebagai X_4, persamaan regresi logistiknya dapat dinyatakan sebagai berikut :

$$\pi(X) = \frac{\exp[g(X)]}{1 + \exp[g(X)]}$$

$$= \frac{\exp\{1{,}045 - 1{,}054[X_1] - 1{,}665[X_2)] - 2{,}450[X_3] - 0{,}447[X_4]\}}{1 + \exp\{1{,}045 - 1{,}054[X_1] - 1{,}665[X_2)] - 2{,}450[X_3] - 0{,}447[X_4]\}}$$

Langkah analisis regresi logistik diulang kembali tetapi dengan hanya memasukkan variabel yang signifikan saja ke dalam langkah analisisnya. Variabel-variabel yang dimasukkan ke dalam persamaan baru semuanya

signifikan terhadap model, hal ini bisa dilihat pada tabel *variable in the equation*(2) pada lampiran 4. Persamaan regresi logistik baru yang didapatkan adalah :

$$\pi(X) = \frac{\exp[g(X)]}{1 + \exp[g(X)]}$$

$$= \frac{\exp\{1{,}641 - 1{,}037[X_1] - 1{,}659[X_2)] - 2{,}422[X_3] - 0{,}448[X_4]\}}{1 + \exp\{1{,}641 - 1{,}027[X_1] - 1{,}659[X_2)] - 2{,}422[X_3] - 0{,}448[X_4]\}}$$

iv) Inferensi dalam regresi logsitik

Setelah estimasi β diketahui, langkah selanjutnya adalah inferensi parameter yang dilakukan dengan uji *Likelihood* rasio

Tabel 3.10 *Omnibus Tests of Model Coefficients*

		Chi-square	df	Sig.
Step 1	Step	212.595	8	.000
	Block	212.595	8	.000
	Model	212.595	8	.000

Pada tabel *Omnibus Tests of Model Coefficients* di atas, nilai Sig pada baris model adalah 0,000. Sig kurang dari α = 0,05 sehingga dapat disimpulkan bahwa model regresi logistik penuh lebih baik daripada model tereduksi (model logistik awal), sehingga persamaan yang dipilih adalah persamaan regresi logistik penuh.

v) Peluang kejadian

Probabilitas event adalah kategori “tidak tepat waktu”, sedangkan kategori non event adalah “tepat waktu”. Dari Tabel 3.10 diketahui bahwa

model regresi logistik biner yang terbentuk mengklasifikasikan sebuah kejadian ke dalam kategori *event* dan *non event* dengan tepat sebesar 64,3%.

Tabel 3.11 Klasifikasi analisis regresi logistik biner

Observed			Predicted		
			Masa Studi		Percentage Correct
			tepat waktu	tidak tepat waktu	
Step 1	Masa Studi	tepat waktu	409	338	54.8
		tidak tepat waktu	233	621	72.7
	Overall Percentage				64.3

vi) Evaluasi Hasil Klasifikasi

$$APER = \frac{233 + 338}{747 + 851} \times 100\% = 35{,}73\ \%$$

$$Press's\ Q = \frac{[N - (nK)]^2}{N(K-1)} = \frac{[1601 - (1029 \times 2)]^2}{1601} = 130{,}449$$

Karena Press's Q lebih besar dari nilai kritiknya (nilai $\chi^2_{(1,0,05)}$) = 3,84 maka ketepatan hasil klasifikasi dapat diterima.

Meskipun nilai APER (*Apparent Error Rate*) tidak sangat rendah, tapi masih dibawah 50%, yaitu 35,73 % sehinggan ketepatan hasil klasifikasi masih dapat diterima (Wuensch, 2009). Jadi dapat disimpulkan bahwa perbedaan yang jelas antara kelompok alumni yang lulus tepat waktu dan tidak tepat waktu ditentukan oleh faktor IP semester 1 dan program studi mereka. Sehingga apabila pihak fakultas ingin meningkatkan kelulusan mahasiswanya, perlu memperhatikan faktor-faktor tersebut.

C. Perbandingan Hasil Klasifikasi Metode CHAID dan Regresi Logistik (biner)

Pada subbab sebelumnya telah diuraikan bagaimana penyelesaian klasifikasi berdasarkan metode CHAID dan analisis regresi logistik biner. Dari segi akurasi, kedua metode tersebut tidak memberikan perbedaan yang cukup besar. Metode CHAID mengklasifikasikan dengan benar sebesar 64,6%, sedangkan analisis regresi logistik biner sebesar 64,3%. Perbedaan yang hanya sebesar 0,3% membuat metode CHAID tidak begitu unggul dalam hal prosentase akurasi, namun selisih sebanyak 0,3% persen tersebut apabila dalam data yang berjumlah ribuan menjadi sangat berpengaruh.

Kelebihan metode CHAID yang tidak dimiliki oleh analisis regresi logistik biner adadlah tentang segmentasi/pengelompokan alumni FMIPA berdasarkan masa studinya. Kalau regresi logistik biner hanya bisa mengelompokkan alumni dengan kategori tepat waktu dan tidak tepat waktu, metode CHAID bisa mengelompokkan lebih dari itu. Metode CHAID dapat mengelompokkan data alumni menjadi beberapa segmen/kelompok dengan kategori dari variabel independen sebagai ciri-ciri dari masing kelompok. seperti pada kasus klasifikasi alumni FMIPA, data yang dianalisis kemudian dikelompokkan mennjadi 8 segmen alumni seperti yang telah disebutkan pada tabel 3.4.

Segmentasi ini membuat tindak lanjut dari hasil klasifikasi menjadi lebih tepat sasaran. Misalnya pada kasus masa studi alumni, telah diketahui segmen yang paling besar kelulusan tidak tepat waktunya adalah alumni yang IP semester 1 kurang dari sama dengan 2,50, program studi non kependidikan dan berasal dari

jalur masuk non regular. Apabila pihak fakultas ingin meminimalisir kelulusan yang tidak tepat waktu pada mahasiswanya, tentu saja mahasiswa yang memiliki ciri-ciri seperti pada segmen alumni yang kelulusan tidak tepat waktunya paling besar harus lebih mendapat perhatian khusus agar jumlah mahasiswa yang lulus tidak tepat waktu bisa dikurangi.

BAB IV

KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan pada pembahasan pada bab sebelumnya dapat disampaikan beberapa kesimpulan sebagai berikut :

1. Secara ringkas, proses klasifikasi dengan metode CHAID terdiri dari beberapa tahap sebagai berikut :
   i. Tahap penggabungan (*merging*), yaitu pemeriksaan tiap variabel independen menggunakan uji independensi *chi-square* untuk menentukan kategori mana yang signifikan untuk menunjukkan perbedaan dalam variabel dependen dan menggabungkan kategori yang tidak signifikan.
   ii. Tahap pemisahan (*splitting*) yaitu pembagian data menggunakan kategori dari variabel independen yang paling signifikan setelah melalui tahap penggabungan.
   iii. Pengulangan tahap *merging* dan *splitting* untuk setiap tingakatan selanjutnya dengan variabel independen sisa yang belum digunakan untuk pemisahan pada tingkatan sebelumnya.
   iv. Pengulangan langkah iii untuk semua subgrup dan hentikan ketika sudah teridentifikasi semua pembagian yang secara statistik telah signifikan (tahap *stopping*).

2. Hasil segmentasi dari Metode CHAID menyebutkan bahwa segmen/kelompok alumni yang kelulusan tidak tepat waktunya paling besar adalah alumni yang IP Semester 1 kurang dari sama dengan 2,50, berasal dari program studi non kependidikan dan jalur masuk non regular. Dan Alumni yang kelulusan tepat waktunya paling besar adalah alumni yang IP semester 1 lebih dari sama dengan 3,51.
3. Perbandingan hasil klasifikasi metode CHAID dan Analisis Regresi Logistik Biner adalah :

Tabel 4.1 Perbandingan hasil klasifikasi metode CHAID dengan regresi logistik biner

Metode CHAID	Analisis Regresi Logistik Biner
Ketepatan hasil klasifikasi 64,6%	Ketepatan hasil klasifikasi 64,3%
Metode CHAID menghasilkan 4 variabel independen yang signifikan terhadap model, yaitu IP semester 1, program studi, jalur masuk dan asal daerah	Analisis regresi logistik biner hanya menghasilkan 2 variabel independen yang signifikan terhadap model, yaitu IP semester 1 dan program studi
Hasil klasifikasi metode CHAID membagi alumni FMIPA menjadi 8 segmen seperti yang tercantum pada tabel 3.4. Segmen yang memiliki kelulusan tepat waktu terbesar adalah alumni dengan IP semester 1 lebih	Pengklasifikasian hanya membagi alumni menjadi 2 kategori *event* dan *non event* yaitu tepat waktu dan tidak tepat waktu.

dari sama dengan 3,51. Segmen alumni yang kelulusan tidak tepat waktunya paling besar adalah alumni dengan IP semester 1 kurang dari 2,50, program studi non kependidikan dan berasal dari jalur masuk nonreguler.	

B. Saran

Setelah membahas klasifikasi dengan metode CHAID, saran yang dapat penulis sampaikan adalah sebagai berikut :

1. Hasil segmentasi bisa digunakan oleh fakultas untuk mengetahui segmentasi mahasiswa FMIPA UNY berdasarkan kemungkinan masa studi yang akan mereka tempuh nantinya
2. Perlu dilakukan penelitian lanjutan serta penambahan variabel independen lain yang lebih menggambarkan latar latar belakang alumni seperti kondisi ekonomi, nilai NEM SMU, nilai tes penerimaan mahasiswa, dll agar akurasi klasifikasi alumni bisa meningkat.
3. Dapat dilakukan penelitian dengan metode serupa pada ruang lingkup yang lain, misalnya pada bidang pendidikan, kesehatan, pemasaran dan perbankan.

DAFTAR PUSTAKA

Ahmad Basuki, Iwan Syarif. (2003). *Decision Tree*. Surabaya: Politeknik Elektronika Negeri Surabaya

Du Toit, S. H. C., A. G. W. Steyn & R. H. Stumph. (1986). *Graphical Exploratory Data Analysis.*New York : Springer-Verlag

Everit, B. S & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics Fourth Edition*. Cambridge : Cambridge University Press

Gallagher, C.A. (2000). *An Iterative Approach to Classification Analysis*. www.casact.org/library/ratemaking/90dp237.pdf. (diakses tanggal 12 Februari 2010)

Haryatmi, S.(1986). *Analisis Data Statistik.* Jakarta : Karunika Universitas Terbuka

Hosmer, D. W. & Lemewshow. (1989). *Applied Logistic Regression.* New York: John Wiley

Kunto, Y.S.dan Hasana, S.N. (2006). *Analisis CHAID Sebagai Alat Bantu Statistika Untuk Segmentasi Pasar*, jurnal Manajemen, Vol. 1 No. 2. Surabaya : Universitas Kristen Petra

Lehmann, T. dan Eherler, D. (2001). *Responder Profiling with CHAID and Dependency Analysis*. www.informatik.unifreiburg.de/~ml/ecmlpkdd/ WSProceedings/w10/lehmann.pdf. (diakses tanggal 20 Maret 2010)

Montgomery, D. C & Peck, E. A. (1992). *Introduction To Linier Regression Analysis Second Edition.* New York : John Wiley&Sons

PBworks. (2007). *Dasar Klasifikasi*. http://intro-dm.pbworks.com/ w/page/20119280/Dasar-Klasifikasi. (diakses tanggal 3 Desember 2010)

Pennington, Ralph.H. (1967). *Introductory Computer Methods and Numerical.* New York : Collier MacMillan

Sheskin, David. (2000). *Handbook of Parametric and Nonparametric Statistical Procedures Second Edition*. Florida : Chapman & Hall.

Wuensch, Karl. (2009). *Binary Logistic with PASW/SPSS*. http://core.ecu.edu/psyc/wuenschk/MV/Multreg/Logistic-SPSS.doc. (diakses tanggal 10 Desember 2010)

Lampiran 1
Deskripsi Data

Frequency Table

Statistics

		Masa Studi	Jenis Kelamin	Asal Daerah	Jalur Masuk	Program Studi	IP Semester 1
N	Valid	1601	1601	1601	1601	1601	1601
	Missing	0	0	0	0	0	0
Mean		1.53	1.75	1.75	1.47	1.46	2.41
Median		2.00	2.00	2.00	1.00	1.00	2.00
Mode		2	2	2	1	1	3
Std. Deviation		.499	.434	.628	.499	.499	.903
Variance		.249	.188	.394	.249	.249	.816
Minimum		1	1	1	1	1	1
Maximum		2	2	3	2	2	4
Sum		2455	2800	2806	2355	2339	3859

Masa Studi

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	tepat waktu	747	46.7	46.7	46.7
	tidak tepat waktu	854	53.3	53.3	100.0
	Total	1601	100.0	100.0	

Jenis Kelamin

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	laki-laki	402	25.1	25.1	25.1
	perempuan	1199	74.9	74.9	100.0
	Total	1601	100.0	100.0	

Asal Daerah

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	DIY	562	35.1	35.1	35.1
	jawa	873	54.5	54.5	89.6
	luar jawa	166	10.4	10.4	100.0
	Total	1601	100.0	100.0	

Jalur Masuk

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	reguler	847	52.9	52.9	52.9
	non reguler	754	47.1	47.1	100.0
	Total	1601	100.0	100.0	

Program Studi

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	pendidikan	863	53.9	53.9	53.9
	non kependidikan	738	46.1	46.1	100.0
	Total	1601	100.0	100.0	

IP Semester 1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<= 2,50	280	17.5	17.5	17.5
	2,51-3,00	563	35.2	35.2	52.7
	3,01-3,50	579	36.2	36.2	88.8
	>= 3,51	179	11.2	11.2	100.0
	Total	1601	100.0	100.0	

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Masa Studi * Jenis Kelamin	1601	100.0%	0	.0%	1601	100.0%
Masa Studi * Asal Daerah	1601	100.0%	0	.0%	1601	100.0%
Masa Studi * Jalur Masuk	1601	100.0%	0	.0%	1601	100.0%
Masa Studi * Program Studi	1601	100.0%	0	.0%	1601	100.0%
Masa Studi * IP Semester 1	1601	100.0%	0	.0%	1601	100.0%

Masa Studi * Jenis Kelamin Crosstabulation

Count

		Jenis Kelamin		Total
		laki-laki	perempuan	
Masa Studi	tepat waktu	172	575	747
	tidak tepat waktu	230	624	854
Total		402	1199	1601

Masa Studi * Asal Daerah Crosstabulation

Count

		Asal Daerah			Total
		DIY	jawa	luar jawa	
Masa Studi	tepat waktu	276	394	77	747
	tidak tepat waktu	286	479	89	854
Total		562	873	166	1601

Masa Studi * Jalur Masuk Crosstabulation

Count

		Jalur Masuk		Total
		reguler	non reguler	
Masa Studi	tepat waktu	438	309	747
	tidak tepat waktu	409	445	854
Total		847	754	1601

Masa Studi * Program Studi Crosstabulation

Count

		Program Studi		Total
		pendidikan	non kependidikan	
Masa Studi	tepat waktu	471	276	747
	tidak tepat waktu	392	462	854
Total		863	738	1601

Masa Studi * IP Semester 1 Crosstabulation

Count

		IP Semester 1				Total
		<= 2,50	2,51-3,00	3,01-3,50	>= 3,51	
Masa Studi	tepat waktu	52	230	331	134	747
	tidak tepat waktu	228	333	248	45	854
Total		280	563	579	179	1601

Lampiran 2
Output Pemrosesan Data dan pembentukan Model dengan CHAID

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	Masa Studi
	Independent Variables	IP Semester 1, Jenis Kelamin, Asal Daerah, Jalur Masuk, Program Studi
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	IP Semester 1, Program Studi, Jalur Masuk, Asal Daerah
	Number of Nodes	13
	Number of Terminal Nodes	8
	Depth	3

Risk

Estimate	Std. Error
.354	.012

Growing Method: CHAID
Dependent Variable: Masa Studi

Classification

	Predicted		
Observed	tepat waktu	tidak tepat waktu	Percent Correct
tepat waktu	501	246	67.1%
tidak tepat waktu	320	534	62.5%
Overall Percentage	51.3%	48.7%	64.6%

Growing Method: CHAID
Dependent Variable: Masa Studi

Lampiran 3
Output Pemrosesan Data dan Pembentukan Model dengan Regresi Logistik Biner

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1601	100.0
	Missing Cases	0	.0
	Total	1601	100.0
Unselected Cases		0	.0
Total		1601	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
tepat waktu	0
tidak tepat waktu	1

Categorical Variables Codings

			Parameter coding		
		Frequency	(1)	(2)	(3)
IP Semester 1	<= 2,50	280	.000	.000	.000
	2,51-3,00	563	1.000	.000	.000
	3,01-3,50	579	.000	1.000	.000
	>= 3,51	179	.000	.000	1.000
Asal Daerah	DIY	562	1.000	.000	
	jawa	873	.000	1.000	
	luar jawa	166	.000	.000	
Jalur Masuk	reguler	847	1.000		
	non reguler	754	.000		
Jenis Kelamin	laki-laki	402	1.000		
	perempuan	1199	.000		
Program Studi	pendidikan	863	1.000		
	non kependidikan	738	.000		

Block 0: Beginning Block

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	2212.301	.134
	2	2212.301	.134

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 2212.301

c. Estimation terminated at iteration number 2 because parameter estimates changed by less than .001.

Classification Table[a,b]

Observed			Predicted		
			Masa Studi		Percentage Correct
			tepat waktu	tidak tepat waktu	
Step 0	Masa Studi	tepat waktu	0	747	.0
		tidak tepat waktu	0	854	100.0
	Overall Percentage				53.3

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.134	.050	7.140	1	.008	1.143

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	ipawal1	179.271	3	.000
		ipawal1(1)	11.760	1	.001
		ipawal1(2)	40.250	1	.000
		ipawal1(3)	64.403	1	.000
		jk(1)	3.234	1	.072
		asal	2.180	2	.336
		asal(1)	2.092	1	.148
		asal(2)	1.798	1	.180
		jalurmasuk(1)	18.455	1	.000
		prodi(1)	47.169	1	.000
	Overall Statistics		199.886	8	.000

Block 1: Method = Enter

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients								
			Constant	ipawal1(1)	ipawal1(2)	ipawal1(3)	jk(1)	asal(1)	asal(2)	jalurmasuk(1)	prodi(1)
Step 1	1	2002.926	-.930	2.112	1.283	.683	.171	.113	.253	-.020	-.398
	2	1999.718	-1.041	2.431	1.393	.781	.200	.134	.292	-.019	-.445
	3	1999.706	-1.045	2.450	1.396	.784	.201	.135	.294	-.019	-.447
	4	1999.706	-1.045	2.450	1.396	.784	.201	.135	.294	-.019	-.447

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 2212.301

d. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	212.595	8	.000
	Block	212.595	8	.000
	Model	212.595	8	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1999.706[a]	.124	.166

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table[a]

Observed			Predicted		
			Masa Studi		Percentage Correct
			tepat waktu	tidak tepat waktu	
Step 1	Masa Studi	tepat waktu	409	338	54.8
		tidak tepat waktu	233	621	72.7
	Overall Percentage				64.3

a. The cut value is .500

Variables in the Equation(2)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ipawal1			124.724	3	.000	
	ipawal1(1)	-1.054	.179	34.528	1	.000	.348
	ipawal1(2)	-1.665	.183	82.551	1	.000	.189
	ipawal1(3)	-2.450	.244	101.193	1	.000	.086
	jk(1)	.201	.124	2.621	1	.105	1.223
	asal			3.565	2	.168	
	asal(1)	.135	.190	.504	1	.478	1.145
	asal(2)	.294	.183	2.582	1	.108	1.341
	jalurmasuk(1)	-.019	.113	.028	1	.868	.981
	prodi(1)	-.447	.109	16.758	1	.000	.640
	Constant	1.405	.222	39.924	1	.000	4.075

a. Variable(s) entered on step 1: ipawal1, jk, asal, jalurmasuk, prodi.

Lampiran 4
Output Analisis Regresi Logistik Biner dengan Hanya Memasukkan Variabel Signifikan Terhadap Model

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1601	100.0
	Missing Cases	0	.0
	Total	1601	100.0
Unselected Cases		0	.0
Total		1601	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
tepat waktu	0
tidak tepat waktu	1

Categorical Variables Codings

		Frequency	Parameter coding		
			(1)	(2)	(3)
IP Semester 1	<= 2,50	280	.000	.000	.000
	2,51-3,00	563	1.000	.000	.000
	3,01-3,50	579	.000	1.000	.000
	>= 3,51	179	.000	.000	1.000
Program Studi	pendidikan	863	1.000		
	non kependidikan	738	.000		

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			Masa Studi		Percentage Correct
			tepat waktu	tidak tepat waktu	
Step 0	Masa Studi	tepat waktu	0	747	.0
		tidak tepat waktu	0	854	100.0
	Overall Percentage				53.3

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.134	.050	7.140	1	.008	1.143

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	prodi(1)	47.169	1	.000
		ipawal1	179.271	3	.000
		ipawal1(1)	11.760	1	.001
		ipawal1(2)	40.250	1	.000
		ipawal1(3)	64.403	1	.000
	Overall Statistics		194.570	4	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	206.415	4	.000
	Block	206.415	4	.000
	Model	206.415	4	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2005.886[a]	.121	.162

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			Masa Studi		Percentage Correct
	Observed		tepat waktu	tidak tepat waktu	
Step 1	Masa Studi	tepat waktu	465	282	62.2
		tidak tepat waktu	293	561	65.7
	Overall Percentage				64.1

a. The cut value is .500

Variables in the Equation(2)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	prodi(1)	-.448	.109	17.003	1	.000	.639
	ipawal1			137.384	3	.000	
	ipawal1(1)	-1.037	.177	34.191	1	.000	.355
	ipawal1(2)	-1.659	.177	87.575	1	.000	.190
	ipawal1(3)	-2.422	.234	107.355	1	.000	.089
	Constant	1.641	.160	105.059	1	.000	5.161

a. Variable(s) entered on step 1: prodi, ipawal1.