Penerapan Metode Regresi Berstruktur Pohon pada Pendugaan Lama Studi Mahasiswa Menggunakan Paket Program R (*Application of Tree Regression in Long Study of University of Jember's Student Use R Program*)

Yuliani Setia Dewi
Staf Pengajar Jurusan Matematika FMIPA Universitas Jember

ABSTRACT

This research aimed to implemate tree regression with one respon and six explanatory variables in R program and apply it to know variables which distinguish long study of University of Jember' student. We can use "tree" function to form tree regression in R. The result of research shows that the rate of long study University of Jember's Student is 1802 days. Based on structure of tree we can know that variables can used to distinguish university of Jember's students in long study are GPA, time used for doing minithesis and faculty From implementation of tree regression,it is known that tree regression can identify variables locally, reveal interactions in the data set, no variable is assumed to follow any kind of statistical distribution and easy to interpret.

Keywords : tree regression, R program , tree function, long study

PENDAHULUAN

Metode regresi berstruktur pohon ini diilhami oleh program AID (Automatic Interaction Detection) yang dikembangkan oleh Morgan dan Sonquist pada tahun 1960-an Dalam bidang statistika, Breiman *et al*., (1984) telah merancang pengambilan keputusan berbasis pohon. Pengantar dalam S tentang model berbasis pohon telah mengembangkan metode dengan fungsi untuk memeriksa dan mengevaluasi pohon serta mempermudah penggunaannya sehingga banyak memberikan dukungan untuk eksplorasi dan analisis data (Clark & Pregibon, 1992).

Berbagai penelitian di bidang lain telah menunjukkan bahwa metode regresi berstruktur pohon merupakan salah satu cara yang menarik dalam melakukan eksplorasi data dan pengambilan kesimpulan, akan tetapi penggunaan model berbasis regresi berstruktur pohon relatif masih belum banyak digunakan. Dari segi komputasi beberapa software tertarik untuk mengembangkan pemodelan regresi berstruktur pohon ini.

Pada kenyataannya, di lapangan cukup banyak mahasiswa yang masa studinya panjang. Dari segi biaya, hal tersebut juga akan menimbulkan masalah. Untuk mengantisipasi hal-hal tersebut di atas, perlu diketahuinya variabel-variabel yang membedakan mahasiswa Universitas Jember dalam hal lama studi.

Dalam penelitian ini, peneliti akan mengimplementasikan program regresi berstruktur pohon di R dengan satu variabel respon dan enam variabel penjelas serta mengaplikasikan metode tersebut untuk mengetahui variabel yang membedakan mahasiswa Universitas Jember dalam hal lama studi serta mengidentifikasi variabel-variabel yang membedakan mahasiswa secara lokal dalam kelompok tertentu (kelompok-kelompok yang terbentuk setelah analisis ini dilakukan), misalnya kelompok Indeks Prestasi Kumulatif (IPK) tertentu, fakultas tertentu, dan sebagainya. Dari hasil implementasi regresi berstruktur pohon diharapkan dapat diketahui ciri-ciri (keunggulan dan kelemahan) dari metode tersebut.

Model Statistika Linier

Model statistika linier merupakan sebuah model yang secara matematis dinyatakan dalam bentuk

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + \varepsilon$$

Dengan *y* adalah variabel acak yang dinamakan respon; x_1, x_2,...,x_k adalah variabel matematis yang nilainya terkontrol (variabel penjelas); ε adalah variabel acak yang menerangkan keragaman acak yang tidak dijelaskan dalam respon dan $\beta_1, \beta_2, \ldots, \beta_k$ adalah konstanta yang nilainya tidak diketahui dan harus diduga dari data.

Model di atas dapat dinyatakan dalam bentuk matriks sebagai

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Penduga kuadrat terkecil dari $\boldsymbol{\beta}$ dinotasikan dengan $\boldsymbol{b}$ dinyatakan oleh

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

dengan $\boldsymbol{X}$ adalah sebuah matriks berukuran n x (k+1) yang berpangkat penuh, $\boldsymbol{\beta}$ adalah vektor parameter tidak diketahui berukuran (k+1)x1 dan $\boldsymbol{\varepsilon}$ adalah vektor acak berukuran n x 1 dengan mean **0** dan varian $\sigma^2\boldsymbol{I}$.

Jika $\boldsymbol{\varepsilon}$ diasumsikan menyebar normal dengan mean **0** dan varian $\sigma^2\boldsymbol{I}$ maka $\boldsymbol{b}$ dan $s^2 = \frac{(\mathbf{y}-\mathbf{bX})'(\mathbf{y}-\mathbf{bX})}{n-p}$ merupakan penduga terbaik diantara semua penduga tak bias lainnya untuk $\boldsymbol{\beta}$ dan σ^2.

Selang kepercayaan 100(1-α)% (penduga selang) untuk parameter β_j, j = 0, 1, 2,..., k dapat dituliskan sebagai berikut

$$b_j \pm t_{\alpha/2} s\sqrt{c_{jj}}$$

dengan $t_{\alpha/2}$ merupakan nilai yang berkaitan dengan sebaran t dengan n-p derajat bebas dan luas di sebelah kanannya $\alpha/2$ dan c_{jj} merupakan konstanta ke-j pada diagonal utama dari matriks $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$. (Myers & Milton, 1991)

Dalam model linier, untuk memperlihatkan interaksi antar variabel dapat dilakukan dengan cara menyertakan suku hasil kali (multiplikatif). Misalnya apabila antara variabel X_1 dan X_2 diperkirakan terdapat interaksi maka modelnya dapat dinyatakan sebagai

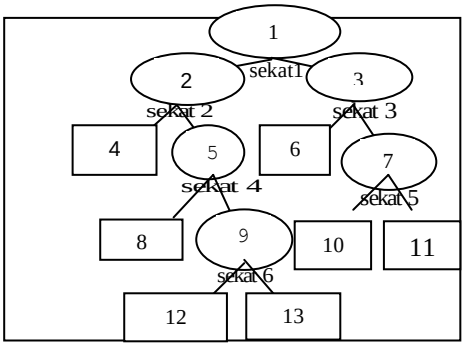
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \underbrace{\beta_3 X_1 X_2}_{\text{sukuinteraksi}} + \ldots + \beta_{k+1} X_k + \varepsilon$$

(Neter *et al.*, 1997)

Regresi Berstruktur Pohon

Regresi berstruktur pohon digunakan untuk mengetahui pengaruh variabel penjelas terhadap variabel respon dan pendugaan respon dilakukan pada kelompok-kelompok pengamatan yang dibentuk berdasarkan variabel-variabel penjelasnya, bukan untuk keseluruhan data.

Metode ini menganalisa suatu gugus data dengan cara menyekatnya menjadi beberapa anak gugus (simpul) secara bertahap sampai tercapai kriteria berhenti tertentu. Anak gugus yang tidak bisa disekat lagi dinamakan simpul akhir (simpul terminal), sedangkan anak gugus yang masih bisa disekat dinamakan simpul dalam. Pada gambar di bawah ini, simpul dalam dilambangkan dengan lingkaran sedangkan simpul akhir dilambangkan dengan persegi panjang.



Gambar 1. Struktur Pohon Regresi

Algoritma Pohon Regresi

Untuk p variabel penjelas X_1, X_2, X_3,...X_p dan satu variabel respon Y yang merupakan variabel numerik, pembentukan pohon regresi memerlukan empat komponen yaitu : Segugus pertanyaan dikotomous Q dengan bentuk "apakah $xi \in A$?" dengan x_i merupakan suatu amatan contoh dan $A \subset X$ (ruang variabel penjelas). Jawaban dari pertanyaan tersebut menentukan sekatan (*partition*) bagi ruang variabel penjelas. Amatan dengan jawaban ya akan masuk ke anak ruang A sedangkan amatan dengan jawaban tidak akan masuk ke anak ruang A^c. Anak ruang yang terbentuk disebut simpul (*node*).

Kriteria kebaikan sekatan (goodness-of – split) $\Phi(s,g)$ merupakan alat evaluasi bagi baik tidaknya penyekatan yang dilakukan oleh sekat s terhadap simpul g.

Aturan penghentian dari proses penyekatan akan menentukan kapan suatu simpul tidak dapat disekat lebih lanjut

Statistik digunakan sebagai ringkasan dari setiap simpul akhir seperti jumlah amatan, nilai dugaan respon, atau besarnya jumlah kuadrat sisaan pada masing-masing simpul akhir.

Tahap Penyekatan

Proses yang dilakukan Breiman *et al* (1993) untuk menyekat suatu simpul adalah sebagai berikut :

Tentukan semua penyekat yang mungkin untuk setiap variabel penjelas. Pilih sekat yang terbaik dari kumpulan sekat dua anak simpul, penyekatan terbaik adalah penyekatan yang memaksimumkan ukuran pemisahan antara dua simpul anak tersebut yakni sekat yang mampu menghasilkan penurunan jumlah kuadrat sisaan terbesar atau $\Phi(s^*,g) = \max_{s^* \in \Omega} \Phi(s,g)$ dengan $\Phi(s,g) = R(g) - R(g_L) - R(g_R)$. Jumlah kuadrat

sisaan di dalam simpul *g* dinyatakan sebagai berikut :

$$R(g) = \sum_{x_n \in g} [y_n - \bar{y}(g)]^2 ,$$

$R(g_L)$ dan $R(g_R)$ adalah jumlah kuadrat sisaan dua simpul anak, yaitu simpul kiri dan kanan.

Pohon regresi dibentuk melalui pemilahan simpul secara rekursif yang memaksimumkan fungsi Φ di atas. Pemilahan tersebut dihentikan tatkala banyaknya amatan dalam simpul tersebut berjumlah "tertentu" atau pada saat nilai Φ lebih kecil dari suatu nilai ambang (*threshold*). Simpul yang terakhir dibentuk disebut sebagai simpul akhir (*terminal node*) atau simpul daun (*leaf node*). Breiman *et al*., (1993) menetapkan banyaknya amatan pada simpul akhir kurang atau sama dengan 5. Suatu simpul akan dijadikan sebagai simpul akhir jika max Φ (s,g) = 0.006JKS(g_1), dengan g_1 adalah simpul utama (Breiman *et al*., 1993). Penentuan nilai ambang bagi Φ analog dengan penentuan α bagi uji-F pada penambahan variabel dalam regresi bertatar langkah maju (*forward stepwise regression*), dimana tak ada lagi penambahan variabel bebas ke dalam model jika variabel tersebut memiliki taraf nyata uji-F yang lebih besar dari α (Therneau & Atiknson, 1997).

Penentuan Ukuran Pohon

Dengan semakin banyaknya jumlah simpul akhir maka struktur pohon akan menjadi semakin kompleks sehingga interpretasi pun sulit dilakukan. Sebaliknya bila simpul akhir berjumlah sedikit, maka struktur pohon yang terbentuk akan semakin sederhana, namun dengan tingkat kesalahan prediksi yang semakin besar. Dengan demikian penentuan ukuran pohon yang layak diperlukan untuk mendapatkan keseimbangan antara tingkat kesalahan prediksi dengan biaya yang muncul akibat kerumitan struktur pohon yang terbentuk.

Suatu upaya yang dapat dilakukan untuk mengatasi hal ini adalah dengan melakukan pemangkasan (*prunning*) terhadap pohon yang terbentuk. Proses ini dimulai dengan membentuk suatu pohon berukuran besar. Pohon yang berukuran besar (G_{max}) ini akan dipangkas menjadi pohon yang lebih kecil. Prosedur pemangkasan dilakukan berdasarkan suatu ukuran biaya kompleksitas (Breiman, *et al*., 1993). Langkah terakhir adalah pemilihan pohon terbaik dari sekuens pohon yang terbentuk. Dalam pemilihan pohon terbaik ini, digunakan suatu penduga yang dinamakan penduga jujur bagi *R(G)* (Breiman et al 1993). Ada dua penduga jujur bagi *R(G)* yaitu penduga contoh uji $R^{ts}(G)$ dan penduga validasi silang $R^{CV}(G)$.

Paket R

R adalah sebuah sistem untuk komputasi dan grafik statistika. Paket R adalah sebuah bahasa yang mirip dengan bahasa S yang dikembangkan di Bell Laboratories oleh John Chambers dan koleganya. R dapat dianggap sebagai sebuah implementasi yang berbeda dari S. Ada beberapa perbedaan penting, tetapi banyak kode yang ditulis untuk bahasa S dapat dijalankan di R. (Venables, 2005)

METODE

Data yang digunakan dalam penelitian ini adalah data lulusan mahasiswa UNEJ bulan Desember 2003 sampai Maret 2005 dengan total respon 2824 orang, dengan variabel tak bebas (respon) nya adalah lama studi mahasiswa (hari) dan variabel bebasnya adalah Jenis kelamin: 1.laki-laki, 2.Perempuan; Fakultas: 1.MIPA, 2.Pertanian, 3 Ekonomi, 4.Hukum, 5.Sastra, 6.FKIP, 7.FKG, 8. FISIP, 9.FTP; lama skripsi (hari); Tempat tinggal : 1.Jember, 2.Luar Jember; Jalur masuk: 1.PMDK, 2.UMPTN dan Indeks Prestasi Kumulatif (IPK). Tahap awal yang dilakukan dalam penelitian ini adalah melakukan telaah pustaka yang berkaitan dengan metode regresi berstruktur pohon. Setelah data terkumpul tahap berikutnya adalah mengimplementasikan program R untuk regresi berstruktur pohon dengan satu variabel respon dan enam variabel penjelas. Tahap selanjutnya adalah melakukan eksplorasi data menggunakan metode regresi berstruktur pohon dengan bantuan paket program R. Kemudian menganalisis hasil yang diperoleh.

HASIL DAN PEMBAHASAN

Implementasi Regresi Pohon Dalam Paket Program R

Paket R telah menyediakan program untuk melakukan analisa data dengan menggunakan regresi pohon. R menyediakan fungsi untuk membentuk, melakukan pemangkasan terhadap pohon yang terbentuk dan juga untuk keperluan validasi silang. Pembentukan pohon regresi di R dapat dilakukan dengan menggunakan fungsi tree. Untuk membentuk regresi berstruktur pohon dengan satu variabel respon (lama studi)

dan enam variabel penjelas (jenis kelamin(JK), fakultas(Fak), jalur masuk (JlrMsk), tempat tinggal (Daerah), lama skripsi dan IPK) dapat dilakukan dengan menjalankan fungsi tersebut.

Pembentukan Pohon

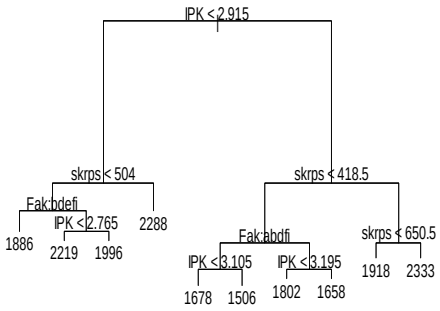
Untuk membentuk pohon regresi dalam R dapat digunakan fungsi tree dan ada metode untuk mencetak (perintah print), merangkum/menampilkan deskripsi data (perintah summary) dan memplot (perintah plot) hasil untuk tree ini. Contoh penggunaan fungsi tree adalah sebagai berikut :

```
>dataunej.tr<-
 tree(lama~JK+Fak+JlrMsk+Daerah+skr
 ps+IPK, dataunej)
> summary(dataunej.tr)
Regression tree:
tree(formula = lama ~ JK + Fak +
JlrMsk + Daerah + skrps + IPK,
    data = dataunej)
Variables actually used in tree
construction:
[1] "IPK"   "skrps" "Fak"
Number of terminal nodes:  10
Residual mean deviance:   78690 =
221400000 / 2814
Distribution of residuals:
Min.     1st  Qu.            Median
Mean        3rd Qu.       Max.
-9.576e+02   -1.683e+02   -4.447e+01
1.924e-14  1.199e+02  2.855e+03

> print(dataunej.tr)

node), split, n, deviance, yval
      * denotes terminal node
 1) root 2824 368900000 1802
   2)  IPK  <  2.915  876  120000000
2035
     4) skrps < 504 722   89350000
1981
       8) Fak: Hukum,KG,KIP,MIPA,TP
411  36540000 1886 *
       9)                      Fak:
Ekonomi,ISIP,Pertanian,SASTRA    311
44250000 2106
        18)   IPK   <   2.765   154
27000000 2219 *
        19)   IPK   >   2.765   157
13380000 1996 *
     5) skrps > 504 154   18690000
2288 *
   3)  IPK  > 2.915  1948  179700000
1697
     6)   skrps   <   418.5   1695
120300000 1652
      12)                      Fak:
Ekonomi,Hukum,KG,MIPA,TP         913
47180000 1577
        24)   IPK   <   3.105   375
24610000 1678 *
        25)   IPK   >   3.105   538
16010000 1506 *
      13)                      Fak:
ISIP,KIP,Pertanian,SASTRA        782
61810000 1740
        26)   IPK   <   3.195   446
37620000 1802 *
        27)   IPK   >   3.195   336
20230000 1658 *
     7) skrps > 418.5 253  33830000
1994
      14)   skrps   <   650.5   207
22310000 1918 *
      15)   skrps   >   650.5   46
5061000 2333 *

>plot(dataunej.tr)
;text(dataunej.tr)
```



Gambar 2. Pohon Regresi untuk Data Lulusan Sarjana Universitas Jember

Model ditentukan dengan formula, komponen-komponennya dipisahkan dengan tanda +. Hasil dari fungsi ini akan menginformasikan kepada kita jika variabel-variabel yang digunakan dalam formula tidak semuanya digunakan untuk membentuk pohon. Dan selanjutnya dengan menggunakan metode print kita mendapatkan printout dari hasil fungsi tree yang berupa banyaknya observasi, jumlah kuadrat (*deviance*) dan nilai rata-rata pada tiap simpul. Perintah plot menghasilkan pohon dengan tidak menggunakan label. Untuk memunculkan label digunakan subperintah text. Secara umum penggunaan fungsi tree dalam R adalah sebagai berikut:

```
tree (formula, data, weights,
subset,
    na.action = na.pass, control =
   tree.control(nobs, ...),
    method = "recursive.partition",
```

```
 split = c("deviance", "gini"),
 model = FALSE, x = FALSE, y =
TRUE, wts = TRUE, ...)
```

Kontrol pada Tree

Ada tiga argumen kontrol pada tree.control, yaitu 'mindev' mengontrol nilai ambang batas untuk dilakukannya penyekatan sebuah simpul. Parameter 'minsize' dan 'mincut' mengontrol ukuran ambang batas. 'Minsize' adalah ambang batas untuk ukuran sebuah simpul,sehingga simpul-simpul dengan ukuran 'minsize' atau lebih besar dapat disekat menjadi simpul anak. Simpul anak harus melebihi 'mincut'. Default dari 'minsize'=10 dan 'mincut'=5. Secara umum penggunaan tree.contol dalam fungsi tree adalah sebagai berikut:

```
tree.control(nobs, mincut = …,
minsize = …., mindev = …..)
```

Pemangkasan Pohon

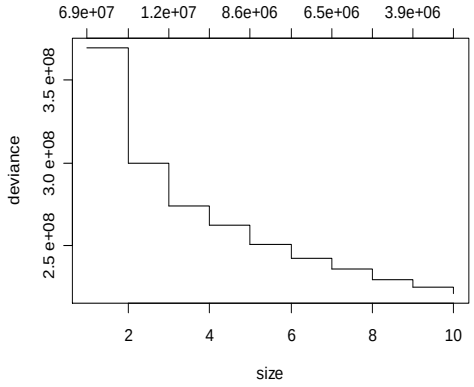
Dengan semakin banyaknya jumlah simpul akhir, struktur pohon yang terbentuk akan semakin rumit, sehingga interpretasi pun sulit dilakukan. Untuk mengatasi hal tersebut dapat dilakukan pemangkasan terhadap pohon.

Dalam R pemangkasan dapat dilakukan dengan menggunakan fungsi prune.tree, yang dapat dilakukan dengan memasukkan satu atau lebih nilai α (melalui argumen k) atau dengan memasukkan suatu ukuran yakni jumlah simpul akhir yang dikehendaki (melalui argument best). Ukuran 'default'nya adalah *deviance.*

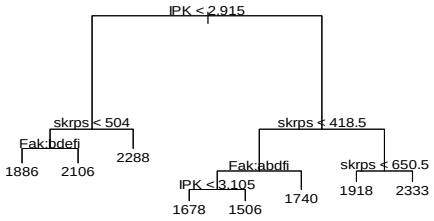
Suatu cara pemilihan α adalah dengan pertimbangan bahwa pemangkasan merupakan sebuah metode penyeleksian variabel, diharapkan mendapatkan nilai Akaike's Information Criterion (AIC) minimum. Untuk regresi pohon, AIC didekati dengan Cp Mallows yakni mengganti -2 x loglikehood dengan jumlah kuadrat sisaan dibagi dengan $\hat{\sigma}^2$, sehingga didapatkan α = 2 $\hat{\sigma}^2$ dengan $\hat{\sigma}^2$ diduga dari model tree penuh. Kriteria lain menyarankan untuk memilih konstanta yang lebih besar dalam wilayah 2 – 6, karena AIC dan Cp cenderung menduga berbias ke atas, (Venables & Ripley, 1994).

Berikut ini contoh penggunaan fungsi prune.tree :

```
> plot(prune.tree(dataunej.tr))
>dataunej.tr1<-
prune.tree(dataunej.tr,best=8)
>
plot(dataunej.tr1);text(dataunej.tr
1)
```



Gambar 3a. *Deviance* dan Jumlah Simpul Akhir dari Pemangkasan Pohon dataunej.tr.

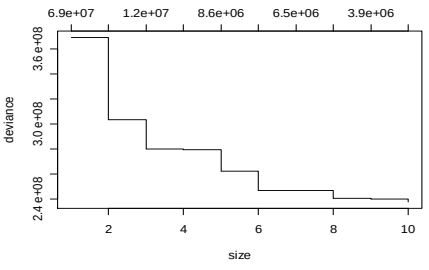


Gambar 3b. Pemangkasan dengan jumlah simpul akhir 8

Validasi Silang

Suatu cara untuk memilih derajat pemangkasan pada paket R adalah dengan menggunakan fungsi cv.tree.

```
>dataunej.cv<-
cv.tree(dataunej.tr,,prune.tree)
> plot(dataunej.cv)
```

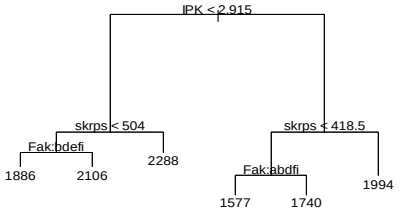


Gambar 4a. Plot Validasi Silang untuk Pemangkasan

Berdasarkan validasi silang yang telah dilakukan kita dapat memilih ukuran pohon

dengan *deviance* mendekati nilai minimum. Misal dari hasil di atas kita pilih pohon dengan ukuran 6, maka dapat kita hasilkan pohon regresi sebagai berikut :

```
>dataunej.tr2<-
prune.tree(dataunej.tr,best=6)
>
plot(dataunej.tr2);text(dataunej.tr
2)
```



Gambar 4b. Pemangkasan dengan Jumlah Simpul Akhir 6

Deskripsi Data

Tabel 1. merupakan gambaran umum mengenai studi mahasiswa Universitas Jember yang berasal dari data wisudawan bulan Desember 2003 sampai Maret 2005 dengan total respon 2824 orang.

Tabel 1. Deskripsi Data Lulusan

Statistik	Lama (hari)	Skripsi (hari)	IPK
Minimum	1144	28,0	2,040
Mean	1802	283,8	3,036
Maksimum	4833	1713,0	3,910

Dari tabel tersebut dapat kita ketahui bahwa mahasiswa Universitas Jember lulus dengan lama studi rata-rata 1802 hari, masa studi paling cepat adalah 1144 hari, sedangkan yang paling lama adalah 4833 hari.

Dalam hal pengerjaan skripsi, rata-rata waktu yang dibutuhkan untuk menyelesaikan skripsi adalah 283,7 hari, skripsi paling cepat adalah 28 hari , dan yang paling lambat adalah 1713 hari.

Mahasiswa Universitas Jember lulus dengan rata-rata IPK 3,036. IPK terkecil lulusan adalah 2,040 dan IPK terbesar adalah 2,910. Deskripsi data untuk masing-masing fakultas adalah sebagai berikut:

Tabel 2. Rata-rata Masa Studi, Masa Skripsi dan IPK Lulusan Unej

Fakultas	Masa Studi	Masa Skripsi	IPK
Ekonomi	1669,641	192,522	3,143
Hukum	1726,884	170,626	3,040
ISIP	1852,054	203,968	3,074
KG	1782,498	346,424	2,872
KIP	1862,151	388,202	3,032
MIPA	1797,384	395,113	2,945
Pertanian	1878,642	314,482	3,033
Sastra	2090,871	385,828	2,995
TP	1665,098	315,390	3,003

Dari Tabel 2 di atas dapat diketahui bahwa rata-rata masa studi yang paling pendek adalah masa studi mahasiswa fakultas TP yaitu 1665,098 hari dan Fakultas Ekonomi yaitu 1669,641 hari sedangkan masa studi yang paling lama adalah masa studi mahasiswa Fakultas Sastra yaitu 2090,871 hari. Dalam hal mengerjakan skripsi, yang lebih cepat dalam mengerjakan skripsi secara rata-rata adalah mahasiswa faktultas hukum (rata-rata 170,626 hari), Fakultas Ekonomi (rata-rata 192,522 hari) dan Fakultas ISIP (203,968 hari) sedangkan yang relatif lama adalah fakultas MIPA (rata-rata 395,113 hari) , Fakultas KIP (rata-rata 388.202 hari) dan Fakultas Sastra (rata-rata 385,828 hari). Dalam hal IPK, Fakultas Ekonomi paling tinggi dibandingkan fakultas-fakultas lainnya.

Pendugaan Lama Studi Mahasiswa Universitas Jember

Berdasarkan hasil di atas, untuk keperluan interpretasi data digunakan hasil pembentukan pohon dengan jumlah simpul akhir 6 (Gambar 4). Dari gambar tersebut dapat dilihat bahwa variabel yang dapat digunakan untuk membedakan mahasiswa Universitas Jember dalam hal lama studinya adalah IPK, lama skripsi dan fakultas. Variabel pertama yang dapat digunakan untuk membedakan adalah IPK. Mahasiswa dengan IPK yang lebih besar, ternyata rata-rata lebih cepat lulus dibandingkan mahasiswa yang IPK-nya lebih kecil. Mahasiswa dengan IPK lebih kecil dari 2,915 rata-rata lama studinya 2035 sedangkan mahasiswa dengan IPK lebih besar dari 2,915 rata-rata lama studinya1697.

Variabel kedua yang membedakan mahasiswa dalam hal lama studi adalah lama skripsi. Semakin lama mahasiswa

menyelesaikan skripsi semakin lama pula mereka lulus. Mahasiswa dengan IPK lebih kecil dari 2,915 : lama skripsi kurang dari 504 hari, lama studinya 1981 hari sedangkan untuk mahasiswa yang lama skripsi lebih dari 504 hari, rata-rata lama studinya 2288 hari. Mahasiswa dengan IPK lebih besar 2,915 : lama skripsi kurang dari 418,5 rata-rata lama studinya 1652 hari, lama skripsi lebih dari 418,5 rata-rata lama studinya 1994 hari.

Variabel ketiga yang membedakan mahasiswa dalam hal lama studi adalah fakultas. Dari hasil pengolahan data diperoleh bahwa untuk mahasiswa dengan IPK lebih kecil dari 2,915 dan lama skripsi kurang dari 504 hari, mahasiswa Fakultas Hukum, KG, KIP, MIPA dan TP (rata-rata lama studi 1886) menyelesaikan studi lebih cepat dibandingkan mahasiswa fakultas lain.(rata-rata lama studi 2106). Sedangkan untuk mahasiswa dengan IPK lebih besar dari 2,915 dan lama skripsi kurang dari 418,5, rata-rata lama studi mahasiswa Fakultas Ekonomi, Hukum, KG, MIPA, TP (rata-rata lama studi 1577) lebih kecil dibandingkan lama studi fakultas lainnya (rata-rata 1740).

Keunggulan dan Kelemahan Regresi Berstruktur Pohon

Dari pembahasan di atas dapat kita ketahui beberapa keunggulan dari regresi berstruktur pohon :

1. Dari segi interpretasi, regresi berstruktur pohon dapat mengidentifikasi variabel-variabel secara lokal, misalnya IPK tertentu, fakultas tertentu dan lainnya.
2. Metode regresi berstruktur pohon tidak memerlukan asumsi variabel harus mengikuti sebaran statistika tertentu
3. Regresi berstruktur pohon dapat mendeteksi dan memperlihatkan interaksi dalam suatu data. Hal ini tidak ditunjukkan oleh model linier baku kecuali ditentukan sebelumnya dengan bentuk multiplikatif tertentu.
4. Skala variabel penjelas dapat berupa campuran antara variabel kategori (misal fakultas) dan kontinyu (misal IPK). Dalam hal ini, interpretasi hasilnya lebih mudah dibandingkan apabila kita menggunakan model linier.

Adapun kelemahan dari metode ini adalah tidak didasarkan pada model peluang. Tidak ada selang kepercayaan yang berkaitan dengan pendugaan yang diturunkan dengan menggunakan regresi berstruktur pohon.

KESIMPULAN

Dari analisis data wisudawan Universitas Jember bulan Desember 2003 sampai Maret 2005 dengan total respon 2824 orang , diperoleh kesimpulan sebagai berikut :

1. Untuk membentuk pohon regresi dalam R dapat digunakan fungsi tree.
2. Dari hasil deskripsi data diketahui bahwa mahasiswa Universitas Jember lulus dengan lama studi rata-rata 1802 hari (4 tahun 11 bulan). Berdasarkan lulusan masing-masing fakultas, rata-rata masa studi yang paling pendek adalah Fakultas TP (rata-rata 1665,098 hari) dan Ekonomi (rata-rata 1669,641 hari) atau 4 tahun 7 bulan sedangkan masa studi yang paling lama adalah Fakultas Sastra yaitu rata-rata 1874,551 hari (5 tahun 2 bulan).
3. Dari hasil struktur pohon yang terbentuk dapat diketahui bahwa variabel yang dapat digunakan untuk membedakan mahasiswa Universitas Jember dalam hal lama studinya adalah IPK, lama skripsi dan fakultas. Mahasiswa dengan IPK yang lebih besar, ternyata rata-rata lebih cepat lulus dibandingkan mahasiswa yang IPK-nya lebih kecil. Semakin lama mahasiswa menyelesaikan skripsi semakin lama pula mereka lulus.
4. Untuk mahasiswa dengan IPK lebih kecil dari 2,915 dan lama skripsi kurang dari 504 hari, lama studi mahasiswa Fakultas Hukum, KG, KIP, MIPA dan TP (rata-rata 1886 hari/5 tahun 2 bulan) menyelesaikan studi lebih cepat dibandingkan mahasiswa fakultas lain(rata-rata 2106 hari/5 tahun 9 bulan). Sedangkan untuk mahasiswa dengan IPK lebih besar dari 2,915 dan lama skripsi kurang dari 418,5, rata-rata lama studi mahasiswa Fakultas Ekonomi, Hukum, KG, MIPA, TP (rata-rata 1577 hari/4 tahun 4 bulan) lebih cepat dibandingkan lama studi fakultas lainnya (rata-rata 1740 hari/4 tahun 9 bulan).
5. Regresi berstruktur pohon dapat mengidentifikasi variabel-variabel secara lokal, tidak memerlukan asumsi sebaran, dapat memperlihatkan interaksi dalam suatu data serta hasilnya mudah diinterpretasikan. Akan tetapi metode ini tidak didasarkan pada model peluang.

DAFTAR PUSTAKA

Breiman L., J.H. Friedman, R.A. Olshen & C.J. Stone, 1984. *Classification and*

Regression Tree. Chapman and Hall, New York

Clark L.A. and Pregibon. 1992. Tree-based models *dalam* Chamber, J.M. & T.J. Hastie (eds). 1992. *Statistical Models in S*. Chapman and Hall, New York.

Venables, 2005. *What is R?*, GNU General Public Lisence. (www.r-project.org/about.html)

Therneau T. M. and E.J. Atkinson, 1997. An Introduction to Recursive Partioning Using the RPART routines. *Technical Report; Mayo Foundation*.

Venables W.N. and B.D. Ripley, 1994. *Modern Applied Statistics with S-Plus*. Springer-Verlag. New York. Inc.