Majalah Ilmiah Matematika dan Statistika Volume 17 Nomor 1 2017, 31 – 40
https://jurnal.unej.ac.id/index.php/MIMS/index ISSN 1411-6669

PENERAPAN METODE REGRESI BERSTRUKTUR POHON PADA PENDUGAAN LAMA PENYUSUNAN SKRIPSI MAHASISWA

(The Application of the Regression Method with Tree Structure on the Estimation of the Student Thesis Duration)

Theresia Trias Candra Dewi, Yuliani Setia Dewi, Agustina Pradjaningsih

Jurusan Matematika, Fakultas MIPA, Universitas Jember

Abstract. The tree regression is one of the regresion methods that can be used to find out the influence of independent X variable to the dependent Y variable. The tree regression method analyzes data by doing step by step isolation. The prunning tree process is used in order to get the optimal tree size. Prunning tree is done based on cost complexity. This paper is written in order to apply the tree regression method in data of graduated student from Faculty of MIPA at 2001-2005 that used to find out the variable which has an influence to the long student minithesis. The research result shows that the long student minithesis is influenced by GPA and department.

Keywords: Clean and Healthy Living Behavior, Health Promotion, Modeling Mathematics.

MSC2020: 62J05

1. Pendahuluan

Salah satu tujuan analisis data dalam statistika adalah untuk mengetahui apakah ada hubungan antara dua variabel atau lebih dan memperkirakan besarnya efek kuantitatif dari perubahan variabel-variabel tersebut. Dalam keperluan analisis data, digunakan analisis regresi untuk mengetahui bentuk hubungan antara dua atau lebih variabel, yaitu sebuah variabel yang akan diramalkan atau disebut juga variabel tak bebas *(dependent variable)* yang dituliskan dalam *Y* dan satu atau lebih variabel yang digunakan untuk meramalkan atau disebut variabel bebas *(independent variable)* yang dituliskan dalam *X*.

Metode regresi pohon merupakan salah satu cara yang menarik dalam melakukan eksplorasi data dan mengambil kesimpulan dalam analisis. Perbedaan regresi berstruktur pohon ini dengan regresi yang biasa digunakan adalah pada regresi berstruktur pohon pendugaan respon dilakukan pada kelompok-kelompok pengamatan yang dibentuk berdasarkan variabel-variabel bebasnya, bukan untuk keseluruhan data.

Metode regresi berstruktur pohon ini menganalisa suatu gugus data dengan cara menyekatnya menjadi beberapa anak gugus (simpul) secara bertahap. Tahap pertama, seluruh data disekat menjadi dua anak gugus kemudian diperiksa kembali secara terpisah dan dibagi lagi berdasarkan penyekat lainnya, demikian seterusnya sampai tercapai kriteria berhenti. Anak gugus yang tidak bisa disekat dinamakan simpul terminal, sedangkan anak gugus yang masih dapat disekat dinamakan simpul dalam.

Permasalahan yang dibahas dalam penelitian ini adalah bagaimana mengaplikasikan metode regresi berstruktur pohon pada data lulusan mahasiswa dan mengidentifikasikan variabel yang berpengaruh terhadap lama penyusunan skripsi mahasiswa.

Regresi berstruktur pohon merupakan salah satu metode yang menggunakan kaidah pohon keputusan yang dibentuk melalui suatu algoritma penyekatan yang dilakukan secara bertahap. Metode regresi berstruktur pohon ini menganalisa suatu gugus data dengan cara menyekatnya menjadi beberapa anak gugus (simpul) secara bertahap. Tahap pertama, seluruh data disekat menjadi dua anak gugus kemudian diperiksa kembali secara terpisah dan dibagi lagi berdasarkan penyekat lainnya, demikian seterusnya sampai tercapai kriteria berhenti. Anak gugus yang tidak bisa disekat dinamakan simpul terminal, sedangkan anak gugus yang masih dapat disekat dinamakan simpul dalam.

Dalam analisis regresi berstruktur pohon diperlukan empat langkah dasar yaitu (Roger L, 2000):

1. proses pembangunan pohon;
2. penghentian proses pembangunan pohon. Dalam proses ini akan ditemukan pohon dengan ukuran yang cukup besar (pohon maksimal);
3. proses pemangkasan (*pruning*) untuk mendapatkan pohon yang cukup sederhana;
4. pemilihan dan pembentukan pohon yang optimal.

Pohon regresi dibentuk melalui penyekatan data pada tiap simpul ke dalam dua simpul anak. Aturannya sebagai berikut :

1. setiap penyekatan tergantung pada nilai yang hanya berasal dari satu variabel bebas;
2. untuk variabel numerik X_j, penyekatan berasal dari pertanyaan "apakah $X_j \le c$" untuk $c \in \Re$. Jadi jika ruang sampelnya berukuran n dan terdapat sebanyak-banyaknya n nilai amatan yang berbeda pada variabel X_j, maka akan terdapat sebanyak-banyaknya n-1 split yang berbeda yang dibentuk oleh gugus pertanyaan ("apakah $X_j \le c_i$"), dengan $i = 1,2,3,...,n-1$ dan c_i adalah nilai tengah antara dua nilai amatan variabel X_j yang berbeda dan berurutan;
3. untuk variabel bebas yang berkategori, pemilahan yang terjadi berasal dari semua kemungkinan pemilahan berdasarkan terbentuknya dua anak gugus yang saling lepas *(disjoint)*. Jika X_j merupakan variabel kategori nominal bertaraf L, maka akan terdapat $2^{L-1}-1$ sekatan yang mungkin, sedangkan jika X_j merupakan variabel kategori ordinal maka akan ada L-1 sekatan yang mungkin.

Menurut Breiman *et al* (1993), untuk menyekat suatu simpul dilakukan proses sebagai berikut :

1. menentukan semua penyekat yang mungkin untuk setiap variabel bebas;
2. memilih sekat yang terbaik dari kumpulan sekat dua anak simpul, yaitu simpul kiri dan simpul kanan.

   Penyekatan terbaik adalah penyekatan yang memaksimumkan ukuran pemisahan antara dua simpul anak tersebut. Jumlah Kuadrat Sisaan (JKS) digunakan sebagai kriteria kehomogenan di dalam masing-masing simpul. Misalkan simpul *g* berisi anak sampel $\{(x_n,y_n)\}$ dan *n(g)* adalah banyaknya amatan pada simpul g, nilai respon dalam suatu simpul g tersebut dapat dihitung sebagai berikut :

   $$\bar{y}(g) = \frac{1}{n(g)} \sum_{x_n \in g} y_n$$

   maka jumlah kuadrat sisaan simpul *g* adalah : $JKS(g) = \sum_{x_n \in g} \left[y_n - \bar{y}(g) \right]^2$

   Misalkan *s* menyekat simpul g menjadi simpul kiri g_L dan simpul kanan g_R. Kriteria jumlah kudrat terkecil $\Phi(s,g)$ adalah :

   $$\Phi(s,g) = R(g) - R(g_L) - R(g_R)$$

   dengan:

   $R(g)$: jumlah kuadrat sisaan pada simpul g atau $JKS(g)$

   $R(g_L)$: jumlah kuadrat sisaan pada simpul kiri g_L atau $JKS(g_L)$

   $R(g_R)$: jumlah kuadrat sisaan pada simpul kanan g_R atau $JKS(g_R)$

   Sekat terbaik s* adalah sekat yang memenuhi kriteria $\Phi(s^*,g) = \max_{s^* \in \Omega} \Phi(s,g)$;
   dengan Ω adalah himpunan semua sekat s yang mungkin pada simpul *g*;
3. algoritma pembentukan struktur pohon dilakukan pada setiap variabel sampai dipenuhi aturan penghentian tertentu. Kriteria yang sering dijadikan aturan penghentian adalah N_{min} banyaknya obyek pengamatan pada setiap simpul akhir;
4. menyusun tingkatan dari semua sekatan terbaik dalam setiap variabel berdasarkan penurunan jumlah kudrat terkecil. Hal ini berarti bahwa sekat yang dipilih untuk dijadikan penyekat utama adalah sekat yang mampu memberikan penurunan jumlah kuadrat sisaan terbesar.

Prosedur pemangkasan dilakukan berdasarkan suatu ukuran biaya kompleksitas (Breiman et al, 1993). Ukuran biaya kompleksitas dari subpohon G_{maks} (pohon berukuran besar atau pohon maksimal), yaitu ukuran biaya dari G, yang didefinisikan sebagai:

$$R_\alpha(G) = R(G) + \alpha \left| \tilde{G} \right|;$$

dengan :

$R_\alpha(G)$: biaya kompleksitas dari G

$|\tilde{G}|$: banyaknya anggota dari gugus simpul akhir $\tilde{G}$

$R(G)$: didefinisikan sebagai $R(G) = \sum_{g' \in \tilde{G}} R(g')$ dengan $R(g')$ adalah jumlah kuadrat sisaan pada suatu simpul akhir g'

α : parameter kompleksitas dengan $\alpha \geq 0$

Pohon yang dipangkas adalah pohon yang memenuhi kriteria biaya kompleksitas minimum.

Dalam memilih pohon terbaik dari deretan pohon yang terbentuk pada proses pemangkasan digunakan suatu penduga yang dinamakan penduga jujur bagi $R(G)$ (Breiman *et al*, 1993). Ada dua penduga jujur bagi $R(G)$ yaitu penduga sampel uji $R^{ts}(G)$ dan penduga validasi silang $R^{CV}(G)$.

$R^{ts}(G_k)$ didefinisikan sebagai:

$$R^{ts}(G_k) = \frac{1}{n_2} \sum_{(x_i, y_i) \in L_2} [y_i - \hat{y}_k(x_i)]^2$$

dengan

$R^{ts}(G_k)$: penduga sampel uji bagi G_k

n_2 : ukuran dari *test sample* L_2

y_i : nilai respon pada amatan ke-i ; $i = 1,2,\ldots,n_2$

$\hat{y}_k(x_i)$: pendugaan respon dari amatan ke-i pada pohon ke-k

Pohon terbaik adalah G_{k0} yang memenuhi : $R^{ts}(G_{k0}) = \min R^{ts}(G_k)$.

Penduga validasi silang $R^{CV}(G_k)$ adalah sebagai berikut:

$$R^{CV}(G_k) = \frac{1}{N} \sum_{v=1}^{v} \sum_{(x_i, y_i) \in L_V} (y_i - \hat{y}_k^v(x_i))^2 ;$$

dengan N adalah jumlah amatan keseluruhan. Pohon terbaik adalah pohon G_{k0} yang memenuhi kriteria:

$$R^{CV}(G_{k0}) = \min R^{CV}(G_k).$$

2. Metodologi

Data yang digunakan dalam penelitian ini adalah data lulusan mahasiswa FMIPA UNEJ periode I bulan November 2001 sampai periode III bulan Maret 2005 dengan total respon 297 orang, dengan variabel tak bebas (respon) nya adalah lama penyusunan skripsi mahasiswa (bulan) dan variabel bebasnya adalah Jenis kelamin: a.laki-laki, b.Perempuan; Jurusan: a.Matematika, b.Fisika, c.Biologi d.Kimia; Asal Daerah: a.Jember, b.Luar Jember; Jalur masuk: a.PMDK, b.UMPTN dan Indeks Prestasi Kumulatif (IPK).

Tahap awal yang dilakukan dalam penelitian ini adalah melakukan telaah pustaka yang berkaitan dengan metode regresi berstruktur pohon. Tahap selanjutnya adalah melakukan eksplorasi data menggunakan metode regresi berstruktur pohon dengan bantuan paket program R. Kemudian menganalisis hasil yang diperoleh.

3. Hasil dan Pembahasan

Berikut ini merupakan gambaran umum mahasiswa FMIPA Universitas Jember yang berasal dari data wisudawan bulan November 2001 sampai Maret 2005 dengan total respon 297 orang.

Tabel 1. Data Lulusan Mahasiswa FMIPA Tahun 2001-2005

	Jenis Kelamin		Asal Daerah		Jalur Masuk	
	L	P	Jember	Luar Jember	PMDK	UMPTN
Matematika	24	44	28	40	14	54
Fisika	20	45	29	36	17	48
Kimia	12	76	50	38	13	75
Biologi	25	51	36	40	9	67
Total	81	216	143	154	53	244

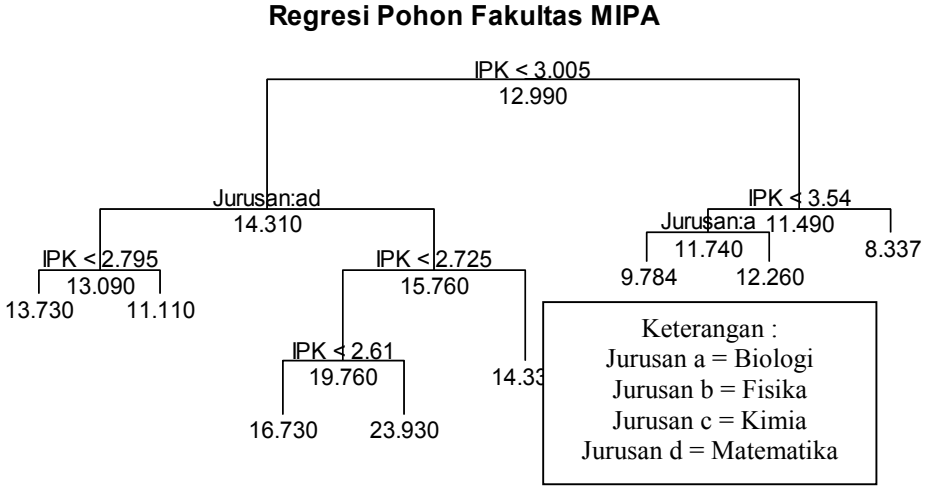
Dari tabel di atas dapat diketahui bahwa mahasiswa perempuan lebih banyak dibandingkan mahasiswa laki-laki. Demikian juga mahasiswa yang masuk melalui jalur UMPTN juga lebih banyak daripada PMDK. Sedangkan jumlah mahasiswa yang berasal dari Jember hampir berimbang dengan yang berasal dari luar Jember. Sedangkan rata-rata IPK dan lama penyusunan skripsi dari keseluruhan mahasiswa (297 sampel) pada masing-masing jurusan adalah sebagai berikut :

Tabel 2. Rata-rata IPK dan Lama Penyusunan Skripsi

	Rata-rata IPK	Rata-rata Lama Penyusunan Skripsi
Matematika	3.04	11.32
Fisika	2.92	11.43
Kimia	2.79	12.77
Biologi	3.04	13.81
FMIPA	2.94	12.99

Formula yang digunakan dalam pembangunan pohon regresi dengan menggunakan paket **R** adalah :

```
mipa.tree<-tree(Lama~JK+Jurusan+Asal+JlrMasuk+IPK,MIPA)
```



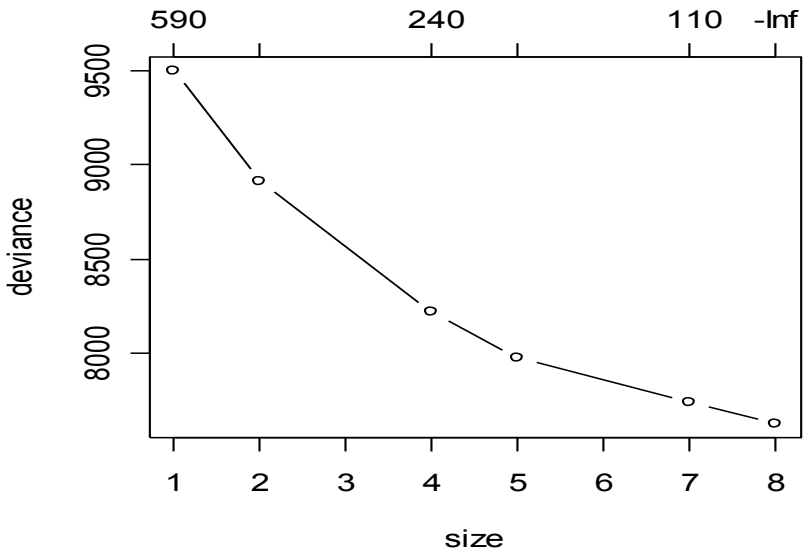
Gambar 1. Pohon Regresi Awal

Berdasarkan proses pembangunan pohon dan hasil plot pohon regresi terlihat bahwa terdapat 8 simpul akhir pada pembangunan pohon awal yaitu :

1. kelompok amatan yang memiliki IPK<2.795 dari jurusan Matematika dan Biologi dengan rata-rata lama penyusunan skripsi 13.73 bulan ;
2. kelompok amatan yang memiliki 2.795<IPK<3.005 dari jurusan Matematika dan Biologi dengan rata-rata lama penyusunan skripsi 11.11 bulan ;
3. kelompok amatan yang memiliki IPK<2.61 dari jurusan Fisika dan Kimia dengan rata-rata lama penyusunan skripsi 16.73 bulan ;
4. kelompok amatan yang memiliki 2.61<IPK<2.725 dari jurusan Fisika dan Kimia dengan rata-rata lama penyusunan skripsi 23.93 bulan ;
5. kelompok amatan yang memiliki 2.725<IPK<3.005 dari jurusan Fisika dan Kimia dengan rata-rata lama penyusunan skripsi 14.33 bulan ;
6. kelompok amatan yang memiliki 3.005<IPK<3.54 dari jurusan Biologi dengan rata-rata lama penyusunan skripsi 9.78 bulan ;
7. kelompok amatan yang memiliki 3.005<IPK<3.54 dari jurusan Matematika, Fisika, dan Kimia dengan rata-rata lama penyusunan skripsi 12.26 bulan ;
8. kelompok amatan yang memiliki IPK>3.54 dengan rata-rata lama penyusunan skripsi 8.34 bulan.

Pemangkasan pohon dilakukan dengan proses *pruning*. Perintah yang dilakukan pada paket **R** :

```
> prune.mipa<-prune.tree(mipa.tree)
```

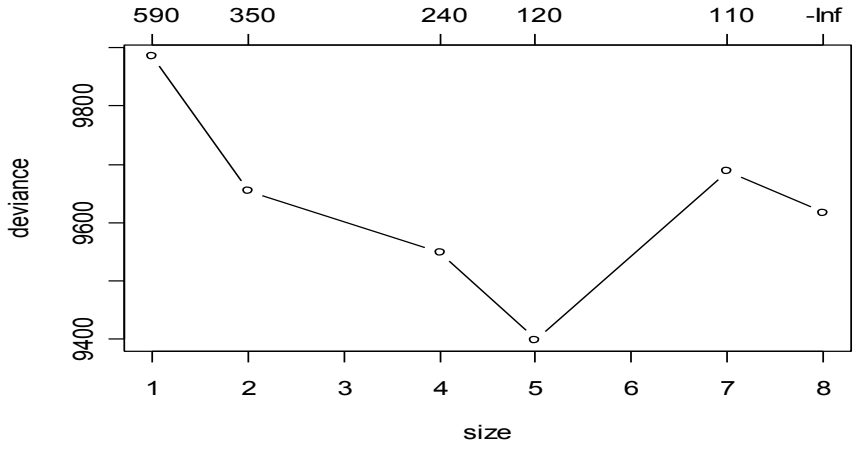


Gambar 2. Hubungan ukuran pohon dan devian dalam pemangkasan

Plot di atas menunjukkan bahwa semakin banyak jumlah pohon yang terbentuk semakin kecil deviannya. Hal ini menunjukkan bahwa biaya kompleksitas yang dibutuhkan untuk pemangkasan akan semakin kecil. Dalam menentukan ukuran pohon digunakan validasi silang *(cross-validation)*. Formula yang digunakan dalam paket R untuk melakukan validasi silang (*cross validation*) adalah cv.tree.

```
> cv.mipa<-cv.tree(mipa.tree)
```

Hasil plot hubungan deretan pohon dan devian berdasarkan validasi silang *(cross validation)* adalah sebagai berikut :

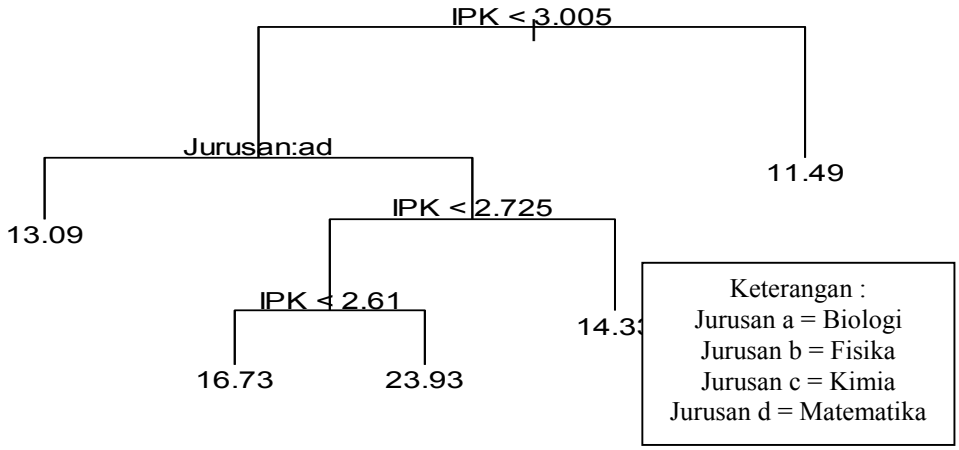


Gambar 3. Hubungan ukuran pohon dan devian berdasarkan validasi silang

Dari hasil plot cv.tree tampak bahwa ukuran pohon yang paling optimal dalam meminimumkan devian adalah simpul dengan ukuran (*size*) 5 yang menunjukkan bahwa pohon optimal memiliki 5 simpul akhir.

Pembentukan pohon optimalyang berukuran 5 adalah sebagai berikut :

```
> prune.mipa<-prune.tree(mipa.tree,best=5)
```



Gambar 4. Plot pohon optimal

Berdasarkan kelompok yang dihasilkan dengan menggunakan analisa regresi pohon diatas dapat disusun urutan kelompok mahasiswa berdasarkan waktu penyusunan skripsi yang lebih cepat sebagai berikut :

1. kelompok mahasiswa dengan waktu penyusunan skripsi paling cepat berada pada simpul 3 yaitu mahasiswa yang memiliki IPK>3,005 dengan rata-rata lama penyusunan skripsi 11,49 bulan;
2. kelompok mahasiswa dengan waktu penyusunan skripsi cepat berada pada simpul 4 yaitu mahasiswa jurusan Matematika dan mahasiswa jurusan Biologi yang memiliki IPK< 3,005 dengan rata-rata lama penyusunan skripsi 13,09 bulan;
3. kelompok mahasiswa dengan waktu penyusunan skripsi sedang berada pada simpul 11 yaitu mahasiswa jurusan Fisika dan mahasiswa jurusan Kimia yang memiliki 2,725<IPK<3,005 dengan rata-rata lama penyusunan skripsi 14,33 bulan;
4. kelompok mahasiswa dengan waktu penyusunan skripsi lama berada pada simpul 20 yaitu mahasiswa jurusan Fisika dan mahasiswa jurusan Kimia yang memiliki IPK<2,61 dengan rata-rata lama penyusunan skripsi 16,73 bulan;
5. kelompok mahasiswa dengan waktu penyusunan skripsi paling lama berada pada simpul 21 yaitu mahasiswa jurusan Fisika dan mahasiswa jurusan Kimia yang memiliki 2,61< IPK<2,725 dengan rata-rata lama penyusunan skripsi 23,93 bulan.

4. Kesimpulan

Dari analisis data wisudawan Universitas Jember bulan November 2001 sampai Maret 2005 dengan total respon 297 orang , diperoleh kesimpulan sebagai berikut :

1. hasil analisa data lulusan mahasiswa FMIPA tahun 2001-2005 menunjukkan bahwa variabel yang paling berpengaruh terhadap lama penyusunan skripsi mahasiswa adalah variabel IPK dan Jurusan ;
2. hasil analisa regresi pohon pada data lulusan mahasiswa FMIPA tahun 2001-2005 berdasarkan IPK, menunjukkan bahwa waktu penyusunan skripsi yang paling cepat terdapat pada kelompok mahasiswa dengan IPK>3,005 dengan rata-rata lama penyusunan skripsi 11,49 bulan ;
3. hasil analisa regresi pohon pada data lulusan mahasiswa FMIPA tahun 2001-2005 berdasarkan jurusan dari kelompok mahasiswa yang memiliki IPK<3,005 menunjukkan mahasiswwa jurusan Matematika dan Biologi memiliki waktu penyelesaian skripsi yang lebih cepat dibandingkan dengan mahasiswa jurusan Fisika dan Kimia.

Daftar Pustaka

[1] Breiman, L. J.H Friedman, R. A. Olshen & Charles J. Stone. (1993), *Classification and Regression Tree*. Chapman & Hall. New York

[2] Chambers J.M & Hastie T.J., (1993), *Statistica Model in S*. Chapman & Hall. New York

[3] Lewis, R.J., (2000), *An Introduction to Classification and Regression Tree (CART) Analysis*. Department of Emergency Medicine Harbor-UCLA Medical Centre. California. http://www.saem.org/download/lewis1.pdf

[4] Venables W.N & Ripley B.D., (1994), *Modern Applied Statistics with S-Plus*. Springer. New York

[5] Yohannes Y & Hoddinot J., (1999), *Classification and Regression Trees : An Introduction*. International Food Policy Research Institute. USA