Forum Statistika dan Komputasi, Oktober 2002, p:19-25 Vol 7 No.2
ISSN 0853-8115

# PENERAPAN METODE PEMANGKASAN DALAM CART (*CLASSIFICATION AND REGRESSION TREE*)

*An Application of prune methods in* CART (*Classification and Regression Tree*)

Taufiq Kurniawan[1], Aunuddin[2], Yayah K. Wagiono[3]

*Abstrak*

*CART (Classification and Regression Tree) is an exploratory data method which analyze the relationship between dependent and independent variables which have large and complex data size i.e.: categorical data (nominal and ordinal) and numerical data (interval and ratio). This method is useful to determine the independent variables. In this study,CART is applied predict the child nutrient.*

*The application of pruning methods in CART is beneficial to avoid over fitting and under fitting in order to make the tree regression. This application uses after treatment child nutrient's data in Kecamatan Bogor Timur. This study shows that the best method is pruning with high decreasing deviance values because it's simple method and can be used for small sample. However, pruning with high decreasing resubstitution relative error and high decreasing cross-validated relative error are more accurate to determine the independent variables also to predict dependent variables with small sample.*

*The application of CART's model to predict child nutrient results consistence and logic independent variables. Therefore CART's model is a precise model to predict child nutrient.*

*Key words : CART Methods*

## PENDAHULUAN

Penggunaan metode berstruktur pohon relatif jarang digunakan oleh statistikawan, meskipun peneliti di bidang lain menemukan beberapa kemudahan dalam penggunaan metode berstruktur pohon yang dapat memberikan lebih banyak informasi dalam pengambilan keputusan (Venables and Ripley, 1994).

Beberapa aplikasi metode pengambilan keputusan berstruktur pohon yaitu studi di bidang botani dan kesehatan, studi di bidang sosial oleh Morgan dan Sonquist tahun 1963 dan Morgan dan Messenger tahun 1973. Pada sekitar tahun 1979, 1983 dan 1986, Quinlan menggunakan metode pengambilan keputusan yang digunakan pada bidang ilmu pembelajaran mesin (*machine learning*) (Venables and Ripley, 1994). Menurut Breiman *et al.* (1984), penggunaan algoritma pohon keputusan dimulai dengan munculnya program *Automatic Interaction Detection* (AID) dari Morgan & Sonquist pada tahun 1963. Metode CHAID *(Chi-square Automatic Interaction Detection)* merupakan teknik eksplorasi untuk menganalisis keterkaitan struktural dalam data kategorik hasil survey dengan menggunakan kriteria khi-kuadrat (Aam Alamudi dkk, 1998). Sedangkan CART (*Classification and Regression Tree*) dipelopori oleh Breiman dan Friedman pada tahun 1973 dan perkembangannya pada tahun 1984 dengan dipeloporinya buku *Classification and Regression Tree* (Abdul Kudus dkk, 1999).

CART memiliki beberapa keunggulan dalam mengeksplorasi struktur data yang komplek, terdiri dari beberapa skala yang berbeda yaitu : nominal, ordinal dan interval dengan jumlah data yang sangat besar. Pertama, peubah tidak bebas dapat berupa data nominal atau ordinal (kategorik) dan interval (skala). Kedua, peubah bebas dapat berupa data nominal, ordinal dan interval. Ketiga, tidak semua peubah bebas diduga dalam level yang sama (nominal, ordinal dan interval). Keempat, data hilang dalam peubah bebas dapat diestimasi dari wakil peubah bebas lainnya. Kelima, hubungan antara banyak variabel sudah diperhitungkan secara simultan untuk dapat menghasilkan model yang terbaik (*www.themeasurementgroup.com*, 2001).

Selain itu, metode ini dapat menentukan struktur data yang lebih sederhana, dan bersifat tegar baik terhadap pencilan maupun asumsi sebaran data. Akan tetapi, kelemahan metode ini

[1]PT ACNielsen Indonesia
[2]Dosen Jurusan Statistika FMIPA IPB
[3]Dosen Jurusan Sosial Ekonomi Pertanian IPB

adalah tidak terukurnya resiko kesalahan α dan besaran dari model keterkaitan antar peubah.

## METODE CART

CART merupakan metode eksplorasi yang digunakan untuk melihat hubungan peubah tidak bebas dengan peubah bebas (*www.themeasurementgroup.com*, 2001). Metode ini dapat mengeksplorasi data yang mempunyai hubungan kausal antara peubah tidak bebas dan peubah bebas dalam jumlah besar yang peubah bebas satu dengan lainnya dapat berinteraksi. Penduga tidak bebas dapat berupa data kategorik (nominal atau ordinal) dan interval.

Metode pohon regresi merupakan metode penyekatan rekursif biner (*binary recursive partitioning*), karena prosesnya simpul (kumpulan data) selalu disekat menjadi dua sekatan yang disebut simpul anak.

Metode pohon regresi dan pohon klasifikasi adalah dua metode yang terkandung dalam metode CART. Kedua metode tersebut berbeda dalam penggunaannya, dimana metode pohon regresi ditujukan untuk pemodelan peubah respon kontinu, sedangkan metode pohon klasifikasi bagi peubah respon kategorik (Abdul Kudus dkk, 1999).

Metode pohon regresi menurut Breiman *et al.* (1984), terdiri dari tiga bagian penting, yaitu :

1. Penyekatan setiap simpul.
2. Penentuan simpul akhir.
3. Penentuan nilai dugaan respon bagi setiap simpul akhir.

### Pemangkasan Pohon Regresi

Bila ukuran pohon tidak terbatas, pohon semakin kompiek dalam menggambarkan data. Pohon regresi yang terlalu besar dapat mengakibatkan *overfitting* dan pohon yang terlalu kecil dapat menimbulkan terjadinya *underfitting* karena tidak terjadinya pemilahan lebih lanjut akibat adanya tetapan nilai ambang $\phi(s^*, t)$ yang sebenarnya pemilahan yang terjadi masih layak. Untuk itu perlu dilakukan suatu teknik yang dapat menyederhanakan ukuran pohon tanpa mengorbankan kebaikan kecocokan model. Teknik ini disebut teknik pemangkasan pada pohon regresi.

Penerapan beberapa metode pemangkasan dalam CART dapat menghasilkan pohon regresi terbaik yang sama baiknya dan memiliki karakteristik yang berbeda-beda. Kriteria pemangkasan tersebut, yaitu :

1. Pemangkasan berdasarkan penurunan nilai devian tertinggi. Pemangkasan yang dilakukan menggunakan ukuran *cost-complexity* minimum. Ukuran *complexity* adalah banyaknya simpul akhir. Dalam regresi biasa, ukuran *complexity* analog dengan derajat bebas model (Therneau & Atkinson dalam Kudus, 1999). Menurut *S-PLUS 2000 Guide to Statistics* (1999), ukuran *cost complexity*, adalah sebagai berikut :

$$D_k(T') = D(T') + k\left|size(T')\right|$$

dimana ,

$D_k(T')$ = devian atau kesalahan pengklasifikasian dari pohon bagian T

$\left|size(T')\right|$ = banyak simpul terminal pada himpunan T′

k = parameter kompleksitas biaya

Pemangkasan parameter kompleksitas biaya menentukan suatu pohon bagian T′ yang meminimumkan $D_k(T')$ pada seluruh pohon bagian. Apabila parameter kompleksitas biaya besar maka banyaknya simpul terminal akan bernilai kecil.

2. Pemangkasan berdasarkan penurunan nilai *resubstitution relative error* (sisaan relatif penggantian ulang) tertinggi. Penentuan nilai sisaan relatif penggantian ulang didapat dari perhitungan *test sample estimate* $R^{ts}(T)$. Menurut Breiman *et al.* (1984), untuk mendapatkan *test sample estimate*, amatan dibagi dua secara acak menjadi *learning sample* $L_1$ dan *test sample* $L_2$. $L_1$ digunakan untuk membentuk urutan pohon $\{T_k\}$ melalui proses pemangkasan, sedangkan $L_2$ digunakan untuk membentuk $R^{ts}(T_k)$. Jika $L_2$ berukuran $n_2$, maka :

$$R^{ts}(T_k) = \frac{1}{n_2} \sum_{(x_n, y_n) \in L_2} \left[ y_n - \hat{y}_k(x_n) \right]^2$$

dimana,

$\hat{y}_k(x_n)$ = dugaan respon dari amatan ke-n pada pohon ke-k.

Pohon terbaik adalah $T_{k0}$, yang memenuhi kriteria :

$$R^{ts}(T_{k0}) = \min_{k} R^{ts}(T_k)$$

3. Pemangkasan berdasarkan selisih nilai tertinggi atau jarak terjauh *cross-validated relative error* (sisaan relatif validasi silang). Menurut Breiman *et al.* (1984), untuk membentuk *cross-validation estimate* dengan *V-fold,* amatan induk L yang berukuran n dibagi secara acak menjadi V kelompok, yakni $L_1, L_2, \ldots, L_v$ yang berukuran sama. *Learning sample* ke-v adalah $L^{-v} = L - L_v$, $v = 1,2, \ldots, V$ yang digunakan untuk membentuk urutan pohon $\{T_k\}$ dan urutan parameter *complexity* $\{\alpha_k\}$. Jadi terdapat v urutan $\{T_k\}$ dan v urutan $\{\alpha_k\}$. Kemudian gunakan amatan induk L untuk membentuk urutan $\{T_k\}$ dan $\{\alpha_k\}$. Definisikan $\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}$ .

   Jika $\hat{y}_k^{-v}(x_n)$ adalah dugaan respon dari amatan ke-n pada pohon yang bersesuaian dengan $\alpha'_k$ yang dibentuk oleh *learning sample* ke-v, maka :

$$R^{CV}(T_k) = \frac{1}{n}\sum_{v=1}^{V} \sum_{(x_n, y_n)\in L_v} \left[ y_n - \hat{y}_k^{-v}(x_n) \right]^2$$

   Pohon terbaik adalah $T_{k0}$, yaitu :

$$R^{CV}(T_{k0}) = \min_{k} R^{CV}(T_k)$$

## PENERAPAN

Data yang digunakan dalam penelitian ini adalah data sekunder hasil penelitian mengenai model pendidikan gizi plus dampaknya terhadap status gizi anak yang dilakukan oleh mahasiswa Doktor Institut Pertanian Bogor pada bulan Februari 2000 sampai Februari 2001 di kecamatan Bogor Timur Jawa Barat.

Jumlah total responden sebesar 131 responden dengan pembagian masing-masing Kelurahan Baranangsiang, Katulampa dan Sukasari yaitu 58, 45 dan 39 responden.

Statistik deskriptif peubah status gizi anak berdasarkan berat badan menurut umur dapat dilihat pada Tabel 1.

Tabel 1. Statistik Deskriptif Peubah Status Gizi Anak

| Peubah | Sesudah Pembinaan |
|---|---|
| Min | -2.73 |
| Max | 1.33 |
| Jangkauan | 4.06 |
| Rataan | -1.0945 |
| Median | -1.18 |
| Q1 | -1.7 |
| Q3 | -0.64 |
| Simp. Baku | 0.8117 |

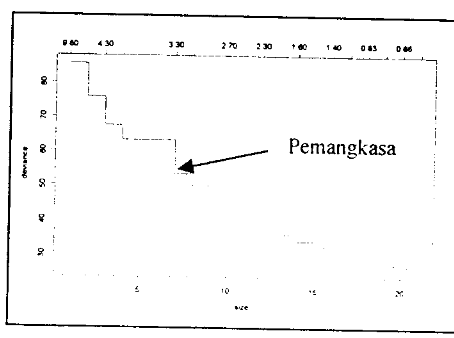
Statistik deskriptif dari peubah - peubah penjelas dapat dilihat pada Lampiran 2. Sedangkan, gambaran data pada peubah lokasi terdiri dari tiga yaitu 1=Kelurahan Baranangsiang, 2=Kelurahan Katulampa dan 3=Kelurahan Sukasari. Pada peubah jenis kelamin dibedakan menjadi dua yaitu 0=laki-laki dan 1=perempuan. Peubah perlakuan terdiri dari dua yaitu 1=tidak memperoleh pembinaan dan 2=memperoleh pembinaan. Sedangkan peubah tipe keluarga terdiri dari dua yaitu 1=inti(ayah, ibu dan anak) dan 2=luas(ayah, ibu, anak, kakek dan saudara lainnya).

### Pohon Regresi Sesudah Pembinaan

Pohon regresi besar menunjukkan peubah status gizi ibu merupakan peubah dominan yang memilah menjadi 2 kelompok besar berdasarkan JKS pohon yang terbesar. Pohon regresi besar ini menghasilkan 21 simpul dengan JKS pohon sebesar 25.99 dan KTS pohon sebesar 0.2362.

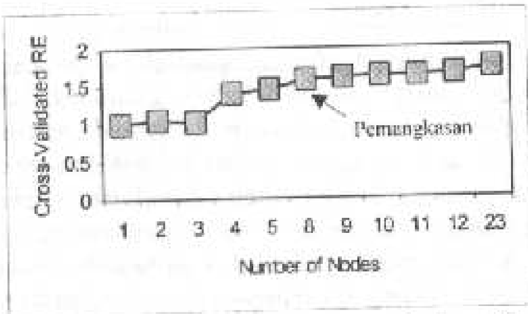
### Pemangkasan

Pemangkasan berdasarkan penurunan nilai devian yang tertinggi ditunjukkan pada simpul ke-7 sebesar 9.96 (Gambar 1) dan menghasilkan nilai devian sebesar 53.25 dengan nilai k (parameter kompleksitas biaya) sebesar 3.29.



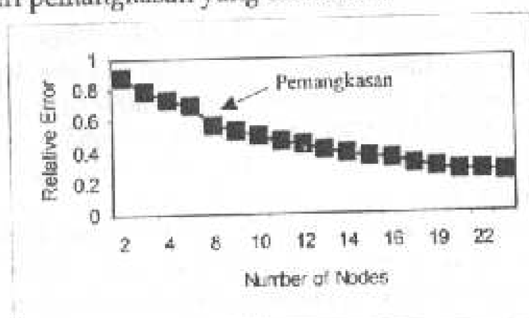
Gambar 1. Grafik Banyak Simpul, Nilai Devian dan Parameter Kompleksitas Biaya (k).

Pemangkasan berdasarkan penurunan sisaan relatif penggantian ulang yang tertinggi ditunjukkan pada simpul ke-8 sebesar 0.123 (Gambar 2) dengan nilai sisaan relatif penggantian ulang 0.573.




Gambar 2. Grafik Banyak Simpul dan Sisaan Relatif Penggantian Ulang

Pemangkasan berdasarkan selisih tertinggi atau jarak sisaan relatif validasi silang terjauh ditunjukkan pada simpul ke-8 sebesar 0.141 (Gambar 3) dengan nilai sisaan relatif validasi silang 1.587 dengan mempertimbangkan makna dari pemangkasan yang dilakukan.




Gambar 3. Grafik Banyak Simpul dan Sisaan Relatif Validasi Silang

Kedua pemangkasan di atas menunjukkan pemangkasan pada simpul ke-8 (Gambar 2 dan 3) dan simpul ke-7 (Gambar 1). Hasil yang didapat dari ketiga metode pemangkasan di atas sebenarnya menghasilkan (output) keluaran yang sama baiknya. Pemangkasan berdasarkan penurunan nilai sisaan relatif penggantian ulang dan sisaan relatif validasi silang lebih akurat dalam penelusuran peubah penyebab. Oleh karena itu pemangkasan dilakukan pada simpul ke-8 dengan JKS pohon 49.95 dan KTS pohon sebesar 0.4061. Pohon regresi terbaik sesudah pembinaan menghasilkan 8 simpul (Lampiran 2) dengan peubah yang paling mempengaruhi status gizi anak yaitu status gizi ibu dan peubah penyebab lainnya yaitu umur ayah, tingkat konsumsi pangan vitamin A pada anak, tingkat konsumsi pangan zat besi pada anak, pengetahuan ibu, alokasi waktu ibu bersama anak dan penyakit diare pada anak.

Kelompok yang terbentuk pada pohon regresi terbaik yaitu sebagai berikut :

1. Tingkat konsumsi pangan zat besi pada anak < 91.15, tingkat konsumsi pangan vitamin A pada anak < 450.35, umur ayah < 32.5 dan status gizi ibu < 22.55 menghasilkan dugaan status gizi anak sebesar –1.38700.
2. Pengetahuan ibu < 2262.5, tingkat konsumsi pangan zat besi pada anak ≥ 91.15, tingkat konsumsi pangan vitamin A pada anak < 450.35, umur ayah < 32.5 dan status gizi ibu < 22.55 menghasilkan dugaan status gizi anak sebesar –1.50100.
3. Pengetahuan ibu ≥ 2262.5, tingkat konsumsi pangan zat besi pada anak ≥ 91.15, tingkat konsumsi pangan vitamin A pada anak < 450.35, umur ayah < 32.5 dan status gizi ibu < 22.55 menghasilkan dugaan status gizi anak sebesar –0.72050.
4. Tingkat konsumsi pangan vitamin A pada anak ≥ 450.35, umur ayah < 32.5 dan status gizi ibu < 22.55 kurang menghasilkan dugaan status gizi anak sebesar –1.83300.
5. Umur ayah ≥ 32.5 dan status gizi ibu < 22.55 menghasilkan dugaan status gizi anak sebesar –0.58670.
6. Penyakit diare pada anak < 240, alokasi waktu ibu bersama anak < 18.75 dan status gizi ibu ≥ 22.55 menghasilkan dugaan status gizi anak sebesar –0.3863.
7. Penyakit diare pada anak ≥ 240, alokasi waktu ibu bersama anak < 18.75 dan status gizi ibu ≥ 22.55 menghasilkan dugaan status gizi anak sebesar 0.6480.
8. Alokasi waktu ibu bersama anak ≥ 18.75 dan status gizi ibu ≥ 22.55 menghasilkan dugaan status gizi anak sebesar –1.05400.

**Perbandingan Metode Pemangkasan**

Pemangkasan pohon regresi sesudah pembinaan berdasarkan penurunan nilai devian tertinggi menghasilkan 7 simpul, penurunan nilai sisaan relatif penggantian ulang tertinggi dan selisih nilai tertinggi atau jarak terjauh antara nilai sisaan relatif validasi silang menghasilkan 8 simpul. Perbedaan yang terjadi karena pada pemangkasan berdasarkan penurunan nilai sisaan relatif penggantian ulang tertinggi dan selisih nilai tertinggi atau jarak terjauh antara nilai sisaan relatif validasi silang lebih akurat dalam penelusuran peubah penyebab.

Kriteria penerapan metode pemangkasan dalam CART memiliki beberapa keunggulan dan kelemahan pada masing-masing metode.

1. Pemangkasan berdasarkan penurunan nilai devian tertinggi mempunyai metode yang lebih praktis dan dapat digunakan dalam ukuran contoh data berukuran kecil
2. Pemangkasan berdasarkan penurunan nilai sisaan relatif penggantian ulang tertinggi dan selisih nilai tertinggi atau jarak terjauh antara nilai sisaan relatif validasi silang lebih akurat dalam penelusuran peubah penyebab dalam menduga peubah respon walaupun dengan ukuran contoh yang lebih kecil. Namun metode pemangkasan sisaan relatif validasi silang akan menghasilkan hasil yang lebih baik apabila dengan ukuran data yang cukup besar. Menurut Breiman *et al.* (1984), ukuran contoh yang dikatakan cukup besar dalam *cross validation estimates* sebesar 900 ukuran contoh dan menurut *CART for windows version 4.0 help* apabila berjumlah lebih besar dari 3000 ukuran contoh

## KESIMPULAN

Penerapan metode pemangkasan dalam CART digunakan untuk menduga status gizi anak dan menghasilkan pemangkasan yang paling baik dengan kriteria, sebagai berikut :

1. Pemangkasan berdasarkan penurunan nilai devian tertinggi karena mempunyai metode yang lebih praktis dan dapat digunakan untuk ukuran contoh kecil.
2. Pemangkasan berdasarkan penurunan nilai sisaan relatif penggantian ulang tertinggi dan selisih nilai tertinggi atau jarak terjauh antara nilai sisaan relatif validasi silang karena lebih akurat dalam penelusuran peubah penyebab dalam menduga peubah respon walaupun ukuran contoh yang lebih kecil

Penerapan model CART untuk menduga status gizi anak menghasilkan peubah penyebab yang konsisten dan logis, oleh karena itu model CART merupakan model yang tepat untuk digunakan dalam pendugaan status gizi anak.

## SARAN

1. Untuk penelusuran peubah penyebab (penjelas) dalam pendugaan status gizi anak dapat menggunakan pendekatan metode analisis yang lain.
2. Peubah penjelas dominan yang dihasilkan CART dapat digunakan untuk analisis selanjutnya.

## DAFTAR PUSTAKA

Alamudi, A., A. H. Wigena & Aunuddin. 1998. Eksplorasi Struktur Data Dengan Metode CHAID. Forum Statistika dan Komputasi. 3.10-16.

Anonym. 1999. *S PLUS 2000 Guide to Statistics Volume I. Data Analysis Product Division MathSoft, Inc.* Seattle, Washington.

Breiman, L., J. H. Friedman, R. A. Olshen & C.J. Stone. 1984. *Classification and Regression Tree.* Chapman and Hall, New York.

Kardiana, A. 1999. Metode Klasifikasi Berstruktur Pohon Biner (Studi Kasus : Prakiraan Sifat Hujan Bulanan di Darmaga Bogor). Thesis. Tidak dipublikasikan.

Kardiana, A., Aunuddin, A. H. Wigena & H Wijayanto. 1999. Penerapan Metode Klasifikasi Berstruktur Pohon Biner Pada Prakiraan Sifat Hujan Harian di Darmaga Bogor. Forum Statistika dan Komputasi. 4:25-30.

Kudus, A. 1999. Penerapan Metode Regresi Berstruktur Pohon pada Pendugaan Masa Rawat Kelahiran Bayi (Studi Kasus di Rumah Sakit Hasan Sadikin Bandung). Thesis. Tidak dipublikasikan.

Kudus, A., Aunuddin, A. H. Wigena & B. Sunarlim. 1999. Eksplorasi Struktur Data Menggunakan Metode Pohon Regresi. Forum Statistika dan Komputasi. 4:12-16.

Nurhayati, A 2000. Faktor-Faktor Yang Berpengaruh Terhadap Status Gizi Anak Usia 6-24 Bulan Di Kecamatan Bandung

Kulon Kotamadya Bandung. Thesis. Tidak dipublikasikan

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

Lampiran 1. Makro SPLUS 2000 Sesudah Pembinaan

**Untuk Menentukan Pohon Regresi Besar Sesudah Pembinaan**

```
status_gizi <- tree(formula = zbbupo ~ tkepo + tkppo + tkcapo + tkphopo + tkfepo + tkvapo +
    tkvcpo + tkbpo + nrkgpo + tptmakpo + rgmmakpo + polakonspo + pengetpo +
    sikappo + homepo + pasuhpo + perl + jeniskel + iumur + aumur + pendi +
    penda + tipekel + besarkel + pdpt + pengel + aset + lingk + fisikibu +
    imt + kepriibu + aksespo + dganakpo + umurpo + ispapo + diarepo +
    ispadiapo + sakitpo + lkpimunpo + tptimunpo, data = data.po.ok,
    na.action = na.exclude, mincut = 5, minsize = 10, mindev = 0.01)
plot(prune.tree(status_gizi))
plot(prune.tree(status_gizi,k=0))
text(prune.tree(status_gizi,k=0))
summary(prune.tree(status_gizi,k=0))
prune.tree(status_gizi,k=0)
```

**Untuk Menentukan Pohon Regresi Terbaik Sesudah Pembinaan**

```
status_gizi <- tree(formula = zbbupo ~ tkepo + tkppo + tkcapo + tkphopo + tkfepo + tkvapo +
    tkvcpo + tkbpo + nrkgpo + tptmakpo + rgmmakpo + polakonspo + pengetpo +
    sikappo + homepo + pasuhpo + perl + jeniskel + iumur + aumur + pendi +
    penda + tipekel + besarkel + pdpt + pengel + aset + lingk + fisikibu +
    imt + kepriibu + aksespo + dganakpo + umurpo + ispapo + diarepo +
    ispadiapo + sakitpo + lkpimunpo + tptimunpo, data = data.po.ok,
    na.action = na.exclude, mincut = 5, minsize = 10, mindev = 0.01)
plot(prune.tree(status_gizi))
plot(prune.tree(status_gizi,k=3.195))
text(prune.tree(status_gizi,k=3.195))
summary(prune.tree(status_gizi,k=3.195))
prune.tree(status_gizi,k=3.19
```

Lampiran 2. Pohon Regresi Terbaik Sesudah Pembinaan