PRAKIRAAN SIFAT HUJAN MENGGUNAKAN METODE POHON KLASIFIKASI

Dwi Haryo Ismunarti

Jurusan Ilmu Kelautan, Fakultas Perikanan dan Ilmu Kelautan, UNDIP

Email: dwiharyois@gmail.com

ABSTRAK

Metode pohon klasifikasi digunakan untuk menduga nilai variabel respon berjenis kategorik, sedangkan variabel bebasnya dapat berjenis kategorik, kontinu atau keduanya. Pohon dibentuk menggunakan algoritma pemilahan secara rekursif terhadap himpunan data pengamatan dan himpunan bagiannya. Pohon klasifikasi untuk prakiraan sifat hujan bulanan menghasilkan Pohon klasifikasi optimum dengan 22 buah simpul terminal dengan nilai harapan tingkat kesalahan pengklasifikasian sebesar 0,33. Variabel Kelembaman nisbi pada jam 13.00 merupakan variabel yang paling berpengaruh. Metode pohon klasifikasi memberikan ketepatan 80% untuk prakiraan sifat hujan.

Kata kunci : pohon klasifikasi. variabel kategorik.

PENDAHULUAN

Eksplorasi data menggunakan metode pohon klasifikasi dikembangkan Breiman *et al* (1984) yaitu *Classification and Regression Trees (CART)*. Metode pohon klasifikasi menggunakan algoritma THAID *(Theta Automatic Interaction Detection)* yang dikembangkan oleh Messenger dan Morgan (Jackson, 1983). Didalam algoritma ini ada 3 tahap pengerjaan berulang untuk memperoleh pohon klasifikasi yang terbaik yaitu : pembentukan pohon, pemangkasan dan pemilihan pohon optimum. Kelebihan penggunaan pohon klasifikasi ini adalah antara lain dapat menangani struktur data yang komplek, interpretasi lebih mudah untuk masing-masing grupnya, identifikasi variabel yang berpengaruh sangat mudah, mempunyai kemampuan untuk mengidentifikasi interaksi antar variabel yang berpengaruh secara lokal sebagai akibat diterapkannya pengambilan keputusan secara bertahap dalam himpunan-himpunan pengukuran.

Dalam suatu penelitian peubah respon tidak selalu merupakan variable terukur yang bersifat kuantitatif. Adakalanya struktur data respon bersifat kualitatif atau kategorik. Nilai peubah kategorik hanya bersifat mengkelaskan observasi ke dalam kelas yang terpisah. Pada penelitian ini pohon klasifikasi akan diterapkan untuk memperkirakan sifat hujan bulanan. Sifat hujan dibedakan atas tiga kategori yaitu Bawah Normal, Normal dan Diatas Normal. Klasifikasi menempatkan sifat hujan ke dalam variabel kategorik. Sedangkan variabel bebasnya merupakan variabel terukur, yaitu: suhu, tingkat penyinaran matahari, tekanan udara, kelembaban nisbi,

indeks osilasi yang merupakan variabel kontinyu.

Metode prakiraan sifat hujan bulanan yang digunakan selama ini adalah regresi linier. Fungsi sebaran variabel yang kontinu yaitu fungsi sebaran normal yang melandasi analisis regresi linear ternyata tidak selalu mencerminkan pola sebaran data yang ada. Tidak terpenuhinya asumsi fungsi sebaran variabel pada metode pendugaan optimum dari analisis regresi linear akan mengakibatkan ketidaktepatan pendugaan (Aunuddin, 1989) dan model yang didapatkan tidak dapat diandalkan (Myers, 1990)

Metode pohon klasifikasi *(Clasification Trees)* dari himpunan data merupakan transformasi monotonik yang akan memilahkan variabel tak bebas y yang berjenis kategorik berdasarkan variabel-variabel bebas x berjenis kategorik, kontinyu ataupun kombinasi keduanya. Berdasarkan jenis variabelnya maka metode pohon klasifikasi dapat diterapkan untuk memperkirakan sifat hujan bulanan. Analisis data digunakan dengan program S Plus 2000.

METODE PENELITIAN

Penerapan metode pohon klasifikasi ini digunakan data sekunder yaitu data klimatologi bulanan yang diamati stasiun klimatologi klas I di Semarang. Data dibagi menjadi 2 kelompok yaitu kelompok data inisialisasi untuk pembentukan model dan kelompok data pengujian (Makridakis dkk, 1988). Kelompok data inisialisasi terdiri dari 240 data klimatologi yaitu 20 tahun x 12 bulan pengamatan. Sedangkan data pengujian terdiri dari 13 data yaitu Nopember 2002 sd Nopember 2003.

Variabel responnya adalah sifat hujan bulanan berjenis kategorik yang nilainya :

a. 0 (bawah normal=BN) jika jumlah curah hujan bulanan < 0,85 x normal curah hujan;
b. 1 (normal =N) jika 0,85 x normal curah hujan ≤ jumlah curah hujan bulanan ≤ 1,15 x normal curah hujan;
c. 2 (atas normal = AN) jika jumlah curah hujan bulanan > 1,15 x normal curah hujan.

Normal curah hujan suatu bulan diperoleh dengan menghitung rata-rata curah hujan bulan tersebut dari mulai data ada sampai tahun sebelum diprakirakan.

Sedangkan variabel penjelasnya yang diamati terdiri dari variabel :

1. Suhu(^{0} C) pada : jam 07.^{00} wib; jam 13.^{00} wib; jam 18.^{00} wib, suhu terbesar; dan suhu terkecil.
2. Tingkat penyinaran matahari dalam %
3. Tekanan udara dalam mb
4. Kelembaban nisbi dalam % pada : jam 07.^{00} wib; jam 13.^{00} wib; jam 18.^{00} wib
5. Sifat hujan bulan sebelumnya
6. Indeks osilasi selatan yaitu beda tekanan udara antara Tahiti dan Darwin dihitung berdasarkan IOS =10 x [dp(Tahiti) – dp(Darwin)]/ sd
   Dp = anomali tekanan udara
   Sd = standar deviasi beda dua anomaly tekanan di Tahiti dan Darwin

Pembentukan pohon klasifikasi meliputi tiga tahapan : Pemilihan pemilah, penentuan simpul dan penandaan kelas. Pemilahan variabel adalah dengan cara memeriksa nilai dari variabel-variabel bebas. Untuk jenis variabel kontinu atau variabel ordinal pemilahan berbentuk

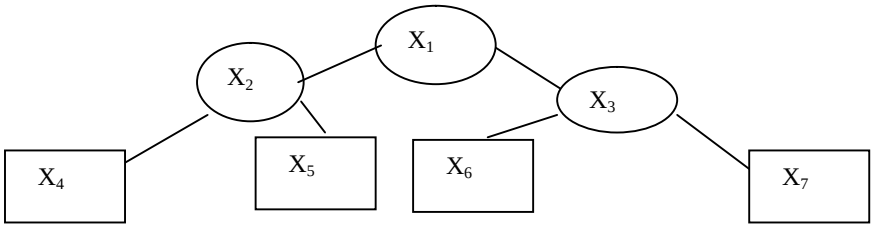
$$x_j \le t \ \ lawan \ \ x_j \ge t; \ \ t \in R.$$

Sedangkan untuk variabel kategorik dengan L taraf akan dibagi menjadi dua himpunan bagian yang saling lepas. Kebaikan pemilah didefinisikan sebagai turunnya keheterogenan yaitu:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

Ketika penurunan nilai keheterogenan dari simpul ke-t tidak lagi berarti atau banyaknya objek terklasifikasi cukup kecil maka pengembangan pohon akan dihentikan dan simpul t ditetapkan sebagai simpul terminal.

Struktur pohon klasifikasi adalah berupa pohon biner yang kemudian akan mempartisi ruang X ke dalam dua himpunan bagian. Dimulai dari X_1 yang dinamakan simpul akar (*root node*) dipilah menjadi X_2 dan X_3 kemudian X_2 dipilah menjadi X_4 dan X_5 sedang X_3 dipilah menjadi X_6 dan X_7 pemilahan akan berhenti/berakhir pada suatu terminal subset atau simpul akhir (*terminal node*) yaitu X_4 , X_5 , X_6 dan X_7 (gambar 1) . Di dalam simpul akhir akan diperoleh suatu nilai yang merupakan prediksi dari variabel respon tersebut.



Gambar 1. Struktur pohon klasifikasi

Misalkan X ruang ukuran yang memuat q vektor atau X ={X_1, X_2, . . . X_q} dan Y himpunan bilangan real. Variabel $X_i \in X$ dinamakan variabel bebas atau variable prediksi dan variabel Y disebut variabel tak bebas atau variable respon. Suatu himpunan data adalah variabel random berdistribusi bersama (X,Y) dengan $X \in R^q$ (Getfard, 1991). Dari himpunan data (X ,Y) akan didefinifikan suatu fungsi bernilai real *d(x)* pada X yang merupakan estimator dari Y.

Definisi 1 :

Fungsi *d(.)* didefinisikan pada X.

d(.): x Č *X* ì *d(x) = y, y* Č*Y.*

Definisi 2:

MSE $R^*(d)$ dari *d* didefinisikan sebagai :

$$R^*(d) = E\,[(Y - d(x)]^2$$

Di dalam pohon klasifikasi $R^*(d)$ merupakan ukuran kesalahan pohon yang disebut *misclassification rate*. Dapat diartikan $R^*(d)$ sebagai harapan kesalahan kuadrat yang dipergunakan *d(x)* sebagai prediksi dari Y (Brieman,et al, 1993) .

Masalah pertama dalam pembentukan pohon klasifikasi adalah bagaimana menggunakan sampel yang ada untuk menentukan pemilah biner sehingga simpul akar *(root node) t* dipilah dalam himpunan bagian-himpunan bagian turunannya. Pemilahan ini akan menyebabkan himpunan bagian tersebut lebih

homogen dibanding dengan induknya (Brieman et al, 1993).

Misalkan suatu pemilah *s* membagi suatu simpul *t* kedalam simpul kiri t_l dan simpul kanan t_R, maka nilai penurunan dari suatu pohon diberikan oleh :

$$\Delta R(s,t) = R(t) - R(t_L) - R(t_R).$$

Kriteria pemilah terbaik s^* diturunkan dari fungsi pemilah *R(s,t)* yang di evaluasi dari pemilah *s* pada suatu simpul *t*. Pemilah terbaik dari suatu simpul *t* adalah pemilah tersebut mempunyai nilai penurunan *R(s , t)* yang terbesar. Sehingga pemilah terbaik adalah yang memaksimumkan nilai dari $\Delta R(s , t)$ tersebut yaitu untuk S himpunan semua calon pemilah:

$$\Delta R(s^*,t) = \max_{s \in S} \Delta R(s,t).$$

Pembentukan pohon dengan kriteria seperti diatas menyebabkan pohon terlalu besar. Permasalahan ini diatasi dengan memangkas *(prune)* pohon tersebut sehingga diperoleh suatu pohon klasifikasi berukuran optimum *(right size tree)*. Suatu ukuran yang dipergunakan di dalam pemangkasan pohon klasifikasi disebut *minimal cost-coplexity pruning* yang didefinisikan sebagai berikut :

Definisi 3:

Untuk suatu pohon bagian $T \leq T_{max}$, didefinisikan *complexity* sebagai $|\overset{\infty}{T}|$ adalah jumlah simpul terminal didalam T. Misal $\alpha \geq 0$ suatu bilangan real adalah parameter *complexity* maka ukuran *cost-complexity* adalah $R_\alpha(T) = R(T) + \alpha |\overset{\infty}{T}|$.

Dengan R(T) adalah ukuran kesalahan pohon T yaitu jumlah kuadrat sisaan dan $|\overset{\infty}{T}|$ adalah ukuran *complexity* yaitu banyaknya simpul terminal. Pemangkasan dimulai secara berturut-turut dari pohon bagian yang kurang penting dengan tingkat kepentingan pohon bagian diukur oleh ukuran *cost-complexity*.

Untuk sembarang T_t yang merupakan cabang dari T_t diberikan ukuran *cost complexity* dari {t} yaitu subcabang dari T_t yang mempunyai satu simpul.

Ukuran cost-complexity subcabang {t} adalah $R_\alpha(\{t\}) = R (t) + \alpha$.

Ukuran cost-complexity dari cabang T_t adalah $R_\alpha(T_t) = R (T_t) + \alpha |\overset{\infty}{T}|$.

Sehingga diperoleh : $R_\alpha(T_t) < R_\alpha(\{ t \})$,

artinya cabang T_t mempunyai *cost complexity* lebih kecil daripada subcabang T_t yang terdiri dari satu simpul { t }.

Suatu nilai kritis dari α diperoleh jika dua ukuran cost complexity tersebut sama yaitu $R_\alpha(T_t) = R_\alpha(\{ t \})$, sehingga nilai α adalah

$$\alpha = \frac{R(t) - R(T_t)}{\left|\overset{\infty}{T_t}\right| - 1}.$$

Didefinisikan suatu fungsi $g_1(t)$. untuk setiap $t \in T_1$.

$$g_1(t) = \begin{cases} \dfrac{R(t) - R(T_t)}{\left|\overset{\infty}{T_t}\right| - 1}, & t \notin \overset{\infty}{T_1} \\ + \infty, & t \in \overset{\infty}{T_1} \end{cases}$$

dengan $\overset{\infty}{T_1}$ adalah himpunan simpul terminal pohon T_1, maka kriteria pemangkasan di dalam T_1 adalah pilih $t \in T_t$ suatu simpul sedemikian sehingga:

$$g_1(t'_1) = \min_{t \in T_1} g_1(t)$$

Algoritma di atas dikerjakan secara berulang sehingga diperoleh barisan pohon yang menurun yaitu : $T_1 > T_2 > \ldots > \{t_1\}$.

Setelah proses pemangkasan tersebut akan diperoleh suatu pohon tersarang dan semakin mengecil $T_1 > T_2 > \ldots > \{t_1\}$ dengan nilai R(T) yang semakin kecil , hal ini menimbulkan masalah sebab dengan kriteria ini akan cenderung memilih pohon yang terbesar dengan nilai R(T) yang terkecil sedang pohon terbesar menyebabkan tidak sederhananya pola data. Untuk mengatasi hal tersebut perlu dipilih pohon yang optimum yaitu pohon yang berukuran sederhana sedemikian hingga nilai R(T) juga cukup kecil dan sudah cukup mampu menggabarkan dari sruktur data yang ada (Brieman et.al.1993). memberikan dua metode estimasi terbaik untuk hal tersebut, yaitu: estimasi sampel uji $R^{ts}(T)$ dan *cross validation V-fold estimate* $R^{CV}(T)$.

Pada estimasi sampel uji himpunan pengamatan L dibagi secara random ke dalam pengamatan L_1 dan L_2. Himpunan data L_1 digunakan untuk pertumbuhan pohon sehingga merupakan barisan pohon $\{T_k\}$ yang merupakan hasil dari pemangkasan. Diambil $d_k(x)$ yang dipergunakan sebagai estimator dari *Y* yang berkorespodensi dengan pohon T_k , apabila L_2 mempunyai N_2 amatan maka nilai dari R(T) dilambangkan dengan $R^{(ts)}(T_k)$ adalah: $R^{(ts)}(T_k) = \frac{1}{N_2} \sum_{(x_n,y_n)\in L_2} (y_n - d(x_n))^2$

dimana pohon yang optimum (T^*) dipilih sedemikian hingga memenuhi kriteria sebagai berikut:

$$R^{ts}(T^*) = \min_k R^{ts}(T_k)$$

Estimasi validasi silang lipat V *(cross validation V-fold)* banyak digunakan dalam keperluan aplikasi dengan nilai estimasinya adalah: $R^{CV}(T_k) = \frac{1}{N} \sum_{1}^{V} \sum_{(Xn,Y_n)\in L_V} (y_n - d^{(V)}{}_k(x_n))^2$

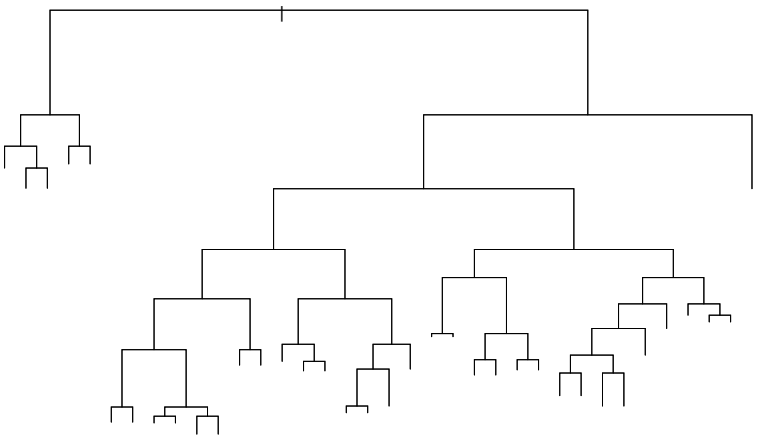
. Pohon optimum (T^*) adalah pohon yang memenuhi kriteria :

$$R^{CV}(T^*) = \min_k R^{CV}(T_k)$$

HASIL DAN PEMBAHASAN

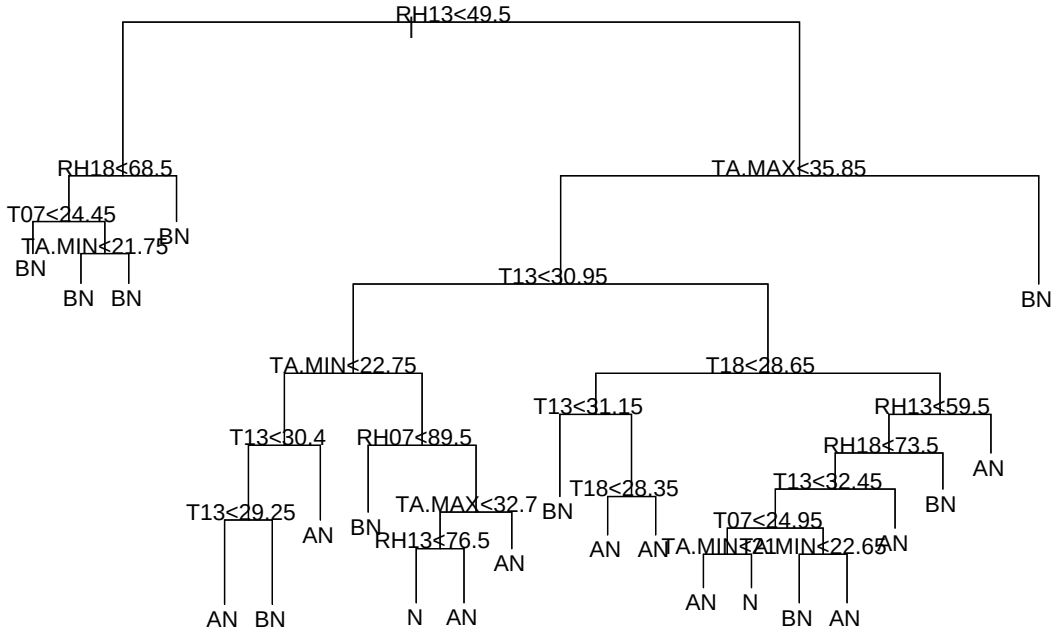
Pohon klasifikasi maksimum diperoleh tardapat pada Gambar 2 dan dapat dijelaskan sebagai berikut:

1. Metode pohon klasifikasi membagi data sebanyak 240 kedalam dua grup (simpul ke-2 dan ke-3) dengan pemilah pertama adalah kelembaban nisbi pada jam 13 sebesar 49.5 % (RH13<49.5) dengan data yang memenuhi sebanyak 43 buah untuk simpul ke-2 dan (RH13>49.5) sebanyak 197 buah untuk simpul yang ke-3. Dengan prakiraan sifat hujan untuk simpul ke-2 adalah dibawah normal (BW) dan prakiraan sifat hujan di atas normal (AN) untuk simpul yang ke-3.
2. Simpul ke-2 membagi data kedalam 2 kelompok yaitu simpul ke-4 dengan jumlah datanya sebanyak 28 buah dan simpul ke-5 dengan data sebanyak 15 . Dengan prakiraan sifat hujan untuk simpul ke-4 adalah dibawah normal (BW) dan prakiraan sifat hujan di atas normal (AN) untuk simpul yang ke-5. Simpul ke-4 ini kemudian dipilah lagi dengan pemilah kelembaban nisbi jam 13 sebesar 68.5 % yaitu RH13<68.5 dan simpul ke-5 dipilah lagi dengan pemilah RH13>68.5.



Gambar 2. Pohon klasifikasi maksimum yang terbentuk

Selanjutnya pohon klasifikasi maksimum yang terbentuk dipangkas secara iteratif menjadi deretan pohon yang makin kecil dan tersarang dengan berdasarkan aturan pemangkasan *cost complexity minimum* (Breimen et al, 1993). Pohon klasifikasi optimum diperoleh dari pemangkasan seperti tampak pada (Gambar 3).



Gambar 3. Pohon Klasifikasi Optimum setelah pemangkasan

Variabel penjelas yang pertama memilah adalah variabel kelembaman nisbi suhu pada jam 13.00 sehingga variabel ini merupakan variabel yang dominan . Pada proses ini terbentuk sebanyak 22 buah simpul terminal dengan kesalahan pengklasifikasiannya sebesar 0,33. Hal ini berarti bahwa pohon klasifikasi yang diperoleh mempunyai ketepatan 67 %.

Pada simpul utama *(root node)* data sebanyak 240 dibagi mejadi dua kelompok kiri dan kanan yang dipilah Juli berdasarkan varibel kelembabab nisbi sebesar 49.5 % (RH13<49.5). Data yang mempunyai rata-rata kelembaman nisbi pada jam 13.00 kurang dari 49.5% sebanyak 43 buah mengelompok di simpul ke-2 dan yang lebih besar dari 49.5% mengelompok di simpul ke-3 sebanyak 197 buah . Pada simpul ke-2 dengan pemilah kelembaman nisbi pada jam 13.00 RH13<49.5 terdiri dari data-data bulan Agustus 1998, September 1998, Juli 1994, Agustus 1994,

Tabel 1. Prakiraan sifat hujan bulan dengan metode Pohon klasifikasi

Th/bln	T07	T13	T18	Tmaks	TMIN	RH07	RH13	RH18	Curah	Sifat	Prakiraan
	^{o}C	^{o}C	^{o}C	^{o}C	^{o}C	%	%	%	Mm	(Xi)	(Fi)
NOV'02	26.4	33.2	28.8	22.7	24.5	81	56	72	272	AN	AN
DES'02	26.7	31.1	28.7	31.8	24.9	84	68	76	148	BN	N *
JAN'03	25.6	30.2	28.2	30.8	24.6	89	72	77	373	N	N
FEB'03	25.4	29.1	26.8	29.8	24.5	90	78	86	568	AN	AN
MRT'03	25.8	30.4	28.1	31.1	24.8	88	72	80	173	BN	BN
APR'03	26.4	31.1	28.6	31.6	24.6	82	63	58	262	AN	AN
MEI'03	24.9	31.1	28.7	31.3	23.9	81	53	68	134	N	AN *
JUN'03	25.4	32.3	29.6	32.9	24.5	79	50	67	0	BN	BN
JUL'03	24	32.6	28.6	33.2	23.1	78	45	69	0	BN	BN
AGS'03	24.4	30.8	27.8	32.3	22.5	72	45	68	0	BN	BN
SEP'03	25.4	31.9	28.2	32.6	23	75	46	66	106	AN	AN
OKT'03	26.2	31.9	28.6	32.7	24.6	80	58	72	264	AN	AN
NOV'03	26.4	31.2	28.3	31.9	24.7	81	64	75	262	AN	AN

- tidak sesuai

Dalam banyak situasi peramalan, ketepatan merupakan kriteria diterima atau ditolaknya suatu model peramalan. Ketepatan menunjukkan seberapa jauh model peramalan mampu mereproduksi data. Bagi pemakai peramalan ketepatan ramalan yang akan datang merupakan hal yang penting. Ukuran ketepatan yang akan digunakan adalah ukuran relatif yaitu nilai tengah kesalahan persentase absolut (Mean Absolute Percentage Error MAPE) (Makridakis,1988).

Tabel 2. Penghitungan MAPE

Th/bln	Sifat hujan	Xi	Prakiraan	Kesalahan		
				Fi	Xi - Fi	|Xi - Fi|/Xi 100
NOV'02	AN	3	AN	3	0	0
DES'02	BN	1	N	2	-1	100
JAN'03	N	2	N	2	0	0
FEB'03	AN	3	AN	3	0	0
MRT'03	BN	1	BN	1	0	0
APR'03	AN	3	AN	3	0	0
MEI'03	N	2	AN	3	-1	50
JUN'03	BN	1	BN	1	0	0
JUL'03	BN	1	BN	1	0	0
AGS'03	BN	1	BN	1	0	0
SEP'03	AN	3	AN	3	0	0
OKT'03	AN	3	AN	3	0	0
NOV'03	AN	3	AN	3	0	0

Catt : BN = 1
N = 2
AN = 3

jumlah 150
MAPE 11.54

Dari tabel 2 di atas tingkat kesalahan peramalan prakiraan sifat hujan bulanan yang dibuat dengan metode klasifikasi adalah sebesar 11.54 % sedangkan tingkat ketepatan peramalan adalah sebesar 88.46 %. Ketepatan metode ini jauh lebih besar dibandingkan ketepatan metode yang digunakan oleh BMG selama ini yaitu dibawah 50 % (Hadiyanto,1994).

KESIMPULAN

Pohon klasifikasi optimum mengandung 22 buah simpul terminal dengan nilai harapan tingkat kesalahan pengklasifikasian sebesar 0,33. Variabel Kelembaman nisbi suhu pada jam 13.00 merupakan variabel yang paling berpengaruh. Metode pohon klasifikasi memberikan ketepatan 88.46% untuk prakiraan sifat hujan pada bulan November 2002 –November 2003 .

DAFTAR PUSTAKA

Aunuddin. 1989. *Analisis Data*. PAU Ilmu Hayat Institut Pertanian Bogor, Bogor.

Breiman, L., et al. 1993. *Classification and Regression Trees*. New York. Chapman & Hall.

Chou, P. A. 1991**.** Optimal Partitioning for Classification and Regression Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 13, No. 4.*

CMyers, R.H. 1990. *Clasical and Modern Regression with Application*. PWS-Kent Publishing Comp: Boston.

Gelfand, S. B., Ravishankar, C. S. dan Delp, E. J. 1991**.** An Iterative Growing and Pruning Classification Tree Design. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, No. 2.*

Hadiyanto , S. 1994 . *Metode Prakiraan Sifat Hujan Bulanan.* Bidang Ramalan dan Jasa, Badan Meteorologi Dan Geofisika Balai Wilayah II (tidak dipublikasikan).

Handoko. 1995. *Klimatologi Dasar*. Ed. Ke-2 Pt Dunia Pustaka Jakarta.

Jackson, B.B. 1983. *Multivariate Data Analysis An Introduction*. Richard D. Irwin , Homewood , Illinois.

Makridakis, S, S.C. Wheelwright dan V.C. MvGee. 1988. *Metode dan Aplikasi Peramalan.* Ed. Terjemahan Andriyanto, U.S. dan A. Basith. Penerbit Erlangga, Jakarta

Miller, A., R.A. Anthes. 1985. *Meteorology*. Abell and Howwel Company Columbus