Klasifikasi Rumah Tangga Miskin di Kabupaten Jombang Berdasarkan Faktor-faktor yang Mempengaruhi dengan Pendekatan CART (*Classification and Regression Trees*)

Riza Inayah, Bambang W. Otok, dan Santi Wulan Purnami
Jurusan Statistika, Fakultas MIPA, Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: santi_wp@statistika.its.ac.id

***Abstrak*—Tingkat kemiskinan di Kabupaten Jombang dapat dikatakan relatif sulit bergerak turun dimana masih ada kurang lebih 73.720 Kepala Keluarga (KK) miskin dari 344 ribu KK yang tersebar di 302 desa dan 4 kelurahan. Pada penelitian ini ingin diperoleh informasi yang akurat mengenai klasifikasi status kemiskinan rumah tangga miskin di Kabupaten Jombang berdasarkan faktor-faktor yang diduga dominan mempengaruhi pengklasifikasian dengan pendekatan CART (*Classification and Regression Trees*). CART termasuk salah satu metode statistik dengan teknik pohon keputusan untuk melakukan analisis klasifikasi dengan pendekatan non-parametrik yang struktur pohonnya diperoleh melalui penerapan prosedur *binary recursive partitioning*. Unit analisis dalam penelitian ini adalah rumah tangga miskin di Kabupaten Jombang dengan jumlah 73.720 rumah tangga. Variabel penelitian meliputi satu variabel respon yaitu status kemiskinan rumah tangga miskin (Rumah Tangga Sangat Miskin atau RTSM, Rumah Tangga Miskin atau RTM, dan Rumah Tangga Hampir Miskin atau RTHM) serta ada delapan belas variabel prediktor. Diperoleh hasil bahwa diantara 73.720 rumah tangga miskin di Kabupaten Jombang tahun 2010, sebanyak 15,8 persen termasuk kelas RTSM, kemudian 39,6 persen termasuk RTM, dan paling banyak yaitu 44,6 persen termasuk kelas RTHM. Variabel terpenting atau paling dominan berpengaruh dalam menentukan status kemiskinan suatu rumah tangga miskin pada penelitian ini yaitu penghasilan rata-rata per bulan (Rp) dengan skor tingkat kepentingan variabel tersebut sebesar 100. Keakuratan hasil klasifikasi oleh penerapan pohon klasifikasi yang optimal untuk data *learning* sebesar 40,986 persen sedangkan data *testing* sebesar 39,654 persen.**

Kata Kunci—Classification and Regression Trees, Binary Recursive Partitioning, Jombang, Klasifikasi Rumah Tangga Miskin

I. PENDAHULUAN

Kemiskinan masih menjadi *tranding topic* bagi pemerintah untuk ditangani dan selalu menjadi perhatian khusus sebab masalah kemiskinan seringkali dijadikan sebagai bahan sorotan untuk mengevaluasi kinerja pemerintah dalam meningkatkan taraf hidup rakyat. Diantara kabupaten/kota yang ada di Jawa Timur, tingkat kemiskinan di Kabupaten Jombang dapat dikatakan relatif sulit bergerak turun. Di Kabupaten Jombang masih ada kurang lebih 73.720 Kepala Keluarga (KK) miskin dari 344 ribu KK yang tersebar di 302 desa dan 4 kelurahan [1]. Berbagai program dan kebijakan baik di bidang sosial, kesehatan, dan sebagainya telah diupayakan untuk menanggulangi masalah tersebut namun belum juga mengatasi kemiskinan yang ada. Sementara itu, oleh [2] ditetapkan kategori status rumah tangga miskin menjadi tiga menurut besarnya pengeluaran per bulan, yaitu Rumah Tangga Sangat Miskin (RTSM), Rumah Tangga Miskin (RTM), dan Rumah Tangga Hampir Miskin (RTHM). Berdasarkan penelitian-penelitian terdahulu yang telah dilakukan, diketahui bahwa banyak faktor yang diduga mempengaruhi tingkat kemiskinan rumah tangga miskin, antara lain terkait kualitas kesehatan, kualitas ekonomi, SDM, dan fasilitas rumah tangga seperti luas lantai, luas kavling bangunan, dan sumber air minum. Namun hasil penelitian tersebut belum cukup mampu memberikan solusi yang membantu pencapaian pengentasan kemiskinan secara menyeluruh dan tepat sasaran untuk masing-masing tingkat kemiskinan rumah tangga di Kabupaten Jombang.

Classification and Regression Trees (CART) merupakan metode statistik dengan pendekatan non-parametrik yang dikembangkan untuk suatu analisis klasifikasi, baik pada variabel respon bertipe kategorik maupun kontinu. CART dapat menyeleksi variabel-variabel prediktor yang paling penting dalam menentukan hasil klasifikasi variabel respon. Dibandingkan dengan metode pengelompokan klasik, CART mempunyai beberapa kelebihan antara lain hasilnya lebih mudah diinterpretasikan karena struktur datanya dapat dilihat secara visual, proses pengklasifikasian lebih mudah dilakukan dengan menelusuri pohon klasifikasi yang dihasilkan, bisa diterapkan untuk data yang kompleks serta tidak memerlukan asumsi tertentu karena CART bersifat non-parametrik sehingga seperti apapun data yang ada bisa lebih fleksibel untuk dianalisis [3].

Ada beberapa penelitian terkait yang telah dilakukan, diantaranya yaitu [4] melakukan analisis metode *ensamble* CART untuk perbaikan klasifikasi kemiskinan di kabupaten Jombang dengan batasan masalah penelitian hanya pada kecamatan Diwek. Hasil analisis menunjukkan bahwa variabel yang digunakan sebagai pemilah pohon klasifikasi CART dan paling menentukan status kemiskinan rumah tangga secara berurutan adalah penghasilan kepala rumah tangga tiap bulan, luas lantai bangunan tempat tinggal, luas

kavling termasuk bangunan, dan sumber air minum. Data sampel yang tepat diklasifikasikan secara keseluruhan sebesar 69,86 persen. Namun, akurasi prediksi pohon klasifikasi CART untuk RTSM hanya 5,02 persen.

Berdasarkan hal-hal tersebut, ada suatu hal penting dan menarik untuk diteliti yaitu tentang informasi yang akurat terkait faktor-faktor yang dominan mempengaruhi klasifikasi tingkat kemiskinan rumah tangga miskin di Kabupaten Jombang. Sehingga nantinya diharapkan dapat berguna dalam strategi perencanaan pembangunan yang lebih terfokus pada pengentasan kemiskinan secara tepat sasaran. Maka dari itu, dalam penelitian ini ingin ditentukan klasifikasi status kemiskinan rumah tangga miskin di kabupaten Jombang tahun 2010 berdasarkan faktor-faktor yang diduga dominan atau paling penting mempengaruhi pengklasifikasian dengan pendekatan CART.

II. TINJAUAN PUSTAKA

A. Classification and Regression Trees (CART)

CART termasuk salah satu metode statistik dengan teknik pohon keputusan untuk melakukan analisis klasifikasi seperti halnya analisis regresi logistik dan analisis diskriminan yang merupakan contoh metode analisis klasifikasi klasik yang sering digunakan. Bedanya,CART menggunakan pendekatan non-parametrik sedangkan analisis regresi logistik dan analisis diskriminan menggunakan pendekatan parametrik. CART akan menghasilkan suatu pohon klasifikasi apabila variabel respon berupa data kategorik, dan akan menghasilkan pohon regresi apabila variabel respon berupa data kontinu. Struktur CART diperoleh melalui penerapan prosedur penyekatan biner yang dapat dilakukan secara berulang kali dalam rangka membentuk partisi-partisi kelas pengamatan yang lebih homogen dengan karakteristik tertentu (*binary recursive partitioning*).

Proses pembentukan pohon klasifikasi membutuhkan data *learning* sehingga sebelumnya perlu dicari dulu metode terbaik untuk pembentukan pohon klasifikasi yaitu yang menghasilkan ketepatan klasifikasi data *testing* tertinggi. Dengan demikian berarti data keseluruhan perlu dibagi menjadi dua himpunan terlebih dulu menjadi L_1 (data *learning*) dan L_2 (data *testing*). Ada dua metode estimasi yang bisa digunakan yaitu estimasi sampel uji (*test sample estimate*) dan estimasi validasi silang lipat-*V* (*Cross Validation V-fold Estimate*). Pada *test sample estimate*, menurut [3] pembagian proporsi L_1 dan L_2 bisa dibagi sesuai ketentuan peneliti namun dianjurkan untuk L_1 jumlahnya lebih banyak dari pada L_2 karena L_1 digunakan dalam membentuk model pohon klasifikasi *T*, sedangkan L_2 digunakan untuk menduga ketepatan klasifikasi data baru oleh pohon klasifikasi *T* yang terbentuk. Sedangkan pada metode *cross validation V-fold* data keseluruhan dibagi menjadi *V* subset yang berukuran relatif sama. Salah satu subset dicadangkan sebagai data *testing* (L_2) dan subset-subset sisanya digabung dijadikan sebagi data *learning* (L_1) dalam prosedur pembentukan pohon klasifikasi. Metode ini akan melakukan pengulangan prosedur pembentukan pohon sebanyak *V* kali dan hasil pengukuran adalah nilai rata-rata dari sebanyak *V* kali pengulangan penerapan prosedur yang dilakukan tersebut. Nilai *V* yang sering dipakai dan dijadikan standar adalah 10 (*cross validation 10-fold*). Sebab hasil dari berbagai percobaandan pembuktian teoritis, menunjukkan bahwa *cross validation 10-fold* adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat.

Pembentukan pohon klasifikasi menggunakan metode estimasi tertentu dengan fungsi keheterogenan tertentu nantinya akan memberikan hasil ketepatan klasifikasi dari data L_1 dan L_2. Jika mencobakan beberapa kombinasi metode estimasi dan fungsi keheterogenan simpul maka akan diperoleh beberapa nilai ketepatan klasifikasi L_1 dan L_2, kemudian dipilih yang terbaik diantara kemungkinan-kemungkinan yang telah dicobakan tersebut yaitu yang menghasilkan ketepatan klasifikasi L_2 tertinggi. Ketepatan klasifikasi L_2 dijadikan sebagai dasar pemilihan metode pembentukan model pohon klasifikasi sebab dapat memberikan gambaran kebaikan pohon klasifikasi yang nantinya terbentuk untuk mengklasifikasikan data baru.

1. Metode Pemilihan Pemilah

Suatu pemilah berasal dari satu kemungkinan pemilah variabel prediktor yang mungkin. Pemilihan pemilah bisa ditentukan menggunakan indeks Gini atau indeks Twoing yang mengukur keheterogenan simpul. Lalu masing-masing variabel dan semua kemungkinan *treshold* atau nilai variabel yang menjadi pemilah dihitung nilai *goodness of split* dan yang menghasilkan nilai maksimum akan dipilih sebagai pemilah terbaik.

Fungsi keheterogenan indeks Gini ditulis pada persamaan (1) berikut.

$$i(t) = \sum_{i \neq j} p(\mathrm{i} \mid \mathrm{t}) p(\mathrm{j} \mid \mathrm{t}) \qquad (1)$$

dengan $i(t)$ merupakan nilai fungsi keheterogenan simpul *t*, $p(i|t)$ merupakan proporsi kelas *i* pada simpul *t*dan $p(j|t)$ merupakan proporsi kelas *j* pada simpul *t.*

Fungsi kehetoreganan indeks Twoing dituliskan pada persamaan (2) berikut.

$$i(t) = \frac{p_L p_R}{4} \left[\sum_j \left| p(j \mid t_L) - p(j \mid t_R) \right| \right]^2 \qquad (2)$$

dengan p_L yaitu proporsi pengamatan pada simpul kiri, p_R yaitu proporsi pengamatan pada simpul kanan. $p(j|t_L)$ yaitu proporsi pengamatan dari simpul *t* menuju simpul kiri dengan kelas *j*, dan $p(j|t_R)$ yaitu proporsi pengamatan dari simpul *t* menuju simpul kanan dengan kelas *j*.

2. Penentuan Simpul Terminal dan Pemberian Label Kelas

Prosedur *binary recursive partitioning*akan berhenti atau dengan kata lain terbentuk simpul terminal yaitu apabila kondisi simpul tersebut memenuhi salah satu kriteria berikut: hanya ada satu pengamatan (n = 1) dalam tiap simpul anak atau adanya batasan minimum *n* pengamatan yang diinginkan peneliti, semua pengamatan dalam setiap simpul anak mempunyai distribusi yang identik terhadap variabel prediktor sehingga tidak mungkin untuk dipilih lagi atau adanya batasan jumlah level atau tingkat kedalaman pohon maksimal yang ditetapkan peneliti. Setiap simpul terminal perlu diberi label kelas sehingga nantinya dapat diketahui karakteristik dari klasifikasi pengamatan untuk setiap kelas variabel respon yang terbentuk. Pemberian label kelas pada simpul terminal

dilakukan berdasarkan aturan jumlah anggota kelas terbanyak pada simpul terminal tersebut.

3. Pemangkasan Pohon (*Pruning*)

Pada tahap pertama sudah terbentuk pohon klasifikasi maksimal (pohon dengan simpul terminal terbanyak).Untuk menghindari terbentuknya pohon klasifikasi yang terlalu besar dan kompleks maka perlu dilakukan proses pemangkasan pohon (*pruning*) agar pohon yang dihasilkan lebih layak. Pemangkasan pohon klasifikasi yaitu suatu penilaian ukuran pohon tanpa mengorbankan ketepatan atau kebaikannya melalui pengurangan simpul pohon yang dianggap tidak begitu signifikan berarti sehingga dicapai ukuran pohon yang layak. Secara berurutan atau bertahap untuk menghasilkan sebuah rangkaian pohon yang sederhana dan lebih sederhana, digunakan metode pemangkasan *cost complexity*. Pada sembarang pohon T yang merupakan sub pohon dari pohon klasifikasi maksimal T_{max} ($T<T_{max}$), dengan nilai $\alpha \geq 0$, maka fungsi *cost complexity* dituliskan dalam persamaan (3).

$$R_a(T) = R(T) + \alpha |\tilde{T}| \quad (3)$$

dengan $R_a(T)$ menunjukkan ukuran kompleksitas suatu pohon T pada kompleksitas α sementara $R(T)$ merupakan ukuran kesalahan klasifikasi pohon *T*. $|\tilde{T}|$ menunjukkan banyaknya simpul terminal pohon *T*. Sedangkan α menunjukkan ukuran kompleksitas oleh penambahan suatu simpul akhir pada pohon *T*. Selanjutnya pencarian pohon bagian $T(\alpha) < T_{max}$ yang meminimumkan $R_a(T)$ yaitu $R_a(T(\alpha)) = \min_{T<T_{max}} R_\alpha(T)$. Suatu pengamatan bisa jadi salah diklasifikasikan [5]. Misal dari contoh Tabel 1 diperoleh informasi bahwa ada sebanyak n_{11} pengamatan kelas 1 yang tepat diklasifikasikan sebagai anggota kelas 1, ada sebanyak n_{12} pengamatan kelas 1 yang salah diklasifikasikan sebagai anggota kelas 2 dan ada sebanyak n_{13} pengamatan kelas 1 yang salah diklasifikasikan sebagai anggota kelas 3. Ukuran keakuratan klasifikasi bisa dihitung sesuai persamaan (4) dimana APER (*Apparent Error Rate*) menunjukkan ukuran kesalahan klasifikasi sejumlah dataset pengamatan oleh suatu fungsi klasifikasi [6].

$$AkurasiKlasifikasi = 1 - APER = \frac{[n_{11} + n_{22} + n_{33}]}{N} \quad (4)$$

Tabel 1.
Contoh Tabel Prediksi Klasifikasi dengan 3Level Kelas Variabel Respon

Kondisi Aktual	Prediksi Klasifikasi Pohon			Total
	Kelas 1	Kelas 2	Kelas 3	
Kelas 1	n_{11}	n_{12}	n_{13}	$n_{1.}$
Kelas 2	n_{21}	n_{22}	n_{23}	$n_{2.}$
Kelas 3	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

B. Teori Kemiskinan

[7] mendefinisikan Garis Kemiskinan (GK) sebagai nilai rupiah yang harus dikeluarkan seseorang dalam sebulan agar dapat memenuhi kebutuhan dasar asupan kalori sebesar 2100 kkal/hari per kapita (Garis Kemiskinan Makanan atau GKM) ditambah kebutuhan minimum non makanan yang merupakan kebutuhan seseorang yaitu papan, sandang, sekolah, transportasi dan kebutuhan individu rumah tangga dasar lainnya (Garis Kemiskinan Non Makanan atau GKNM).

[8] menetapkan kategori status kemiskinan rumah tangga miskin sebagai berikut.

(i) Golongan Rumah Tangga Sangat Miskin (RTSM): rumah tangga yang mengkonsumsi makanan senilai sampai dengan 1.900 kalori per hari, yang senilai dengan Rp 120.000,- per minggu atau bila disetarakan dengan pengeluaran per bulan adalah Rp 480.000,- per rumah tangga per bulan.

(ii) Golongan Rumah Tangga Miskin (RTM): rumah tangga yang mengkonsumsi makanan senilai sampai 2.100 kalori per hari, yang senilai dengan Rp 150.000,- per minggu atau bila disetarakan dengan pengeluaran per bulan adalah Rp 600.000,- per rumah tangga per bulan.

(iii) Golongan Rumah Tangga Hampir Miskin (RTHM): rumah tangga yang mengkonsumsi makanan senilai sampai dengan 2.300 kalori per hari, yang senilai sampai dengan Rp 175.000,- per minggu atau bila disetarakan dengan pengeluaran per bulannya adalah Rp 700.000,- per rumah tangga per bulan.

[2] menggunakan empat belas indikator kemiskinan untuk memenuhi berbagai program pelayanan dasar data rumah tangga yang meliputi luas lantai rumah, jenis lantai rumah, jenis dinding rumah, fasilitas tempat buang air besar, sumber air minum, penerangan yang digunakan, bahan bakar yang digunakan, frekuensi makan dalam sehari, kebiasaan membeli daging/ayam/susu, kemampuan membeli satu set pakaian, kemampuan berobat ke puskesmas/ poliklinik, lapangan pekerjaan kepala rumah tangga, pendidikan kepala rumah tangga, dan kepemilikan asset.

III. METODOLOGI PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data sekunder dari hasil Survei Verifikasi Rumah Tangga Miskin di Kabupaten Jombang yang dirancang oleh Badan Perencanaan Pembangunan (Bappeda) Kabupaten Jombang tahun 2010. Unit analisis dalam penelitian ini adalah rumah tangga miskin di Kabupaten Jombang yang jumlahnya mencapai 73.720 rumah tangga.

Tabel 2.
Variabel Penelitian

	Variabel	Skala Data	Kategori
Y	Status kemiskinan rumah tangga miskin menurut BPS	O	1. Rumah Tangga Sangat Miskin (RTSM) 2. Rumah Tangga Miskin (RTM) 3. Rumah Tangga Hampir Miskin (RTHM)
X_1	Status penguasaan bangunan tempat tinggal	N	1. Milik sendiri 2. Kontrak 3. Sewa 4. Bebas sewa 5. Rumah dinas 6. Rumah milik orang tua/ sanak saudara
X_2	Luas kavling termasuk bangunan (m^2)	R	-
X_3	Luas lantai (m^2)	R	-
X_4	Jenis atap rumah terluas	N	1. Beton 2. Genteng 3. Kayu Sirap 4. Seng 5. Asbes 6. Ijuk/rumbia

Tabel 2. Lanjutan

	Variabel	Skala Data	Kategori
X_5	Jenis dinding terluas	N	1. Tembok 2. Kayu 3. Bambu
X_6	Jenis lantai terluas	N	1. Keramik/ marmer 2. Ubin/ tegel 3. Semen/ bata merah 4. Kayu/papan 5. Bambu 6. Tanah
X_7	Fasilitas tempat buang air besar (jamban)	N	1. Milik sendiri 2. Milik bersama 3. Umum 4. Tidak ada
X_8	Tempat pembuangan akhir tinja	N	1. Septictank 2. Kolam/sawah 3. Sungai/waduk 4. Lubang tanah 5. Tanah lapang/ kebun
X_9	Sumber penerangan utama	N	1. Listrik PLN meteran 2. Listrik PLN bukan meteran (menumpang, dsb) 3. Listrik Non PLN 4. Bukan Listrik
X_{10}	Sumber air minum	N	1. Air dalam kemasan 2. Ledeng 3. Pompa 4. Sumur 5. Mata air 6. Air sungai
X_{11}	Bahan bakar memasak	N	1. Listrik 2. Gas/elpiji 3. Minyak tanah 4. Arang kayu/ tempurung 5. Kayu bakar
X_{12}	Intensitas konsumsi daging/ susu/ ayam per minggu	R	-
X_{13}	Intensitas membeli pakaian per tahun	R	-
X_{14}	Intensitas makan per hari	R	-
X_{15}	Pengobatan	N	1. RS/ Puskesmas/ Pustu 2. Praktik dokter 3. Praktik paramedis 4. Praktik pengobatan tradisional
X_{16}	Ijazah terakhir kepala keluarga	N	1. Tidak punya 2. SD/ setara 3. SLTP/ setara 4. SLTA/ setara 5. Diploma I/ II 6. Akademi ke atas
X_{17}	Penghasilan tiap bulan (Rp)	R	-
X_{18}	Kepemilikan aset (Rp)	R	-

ket: N= Nominal, O = Ordinal, R=Rasio

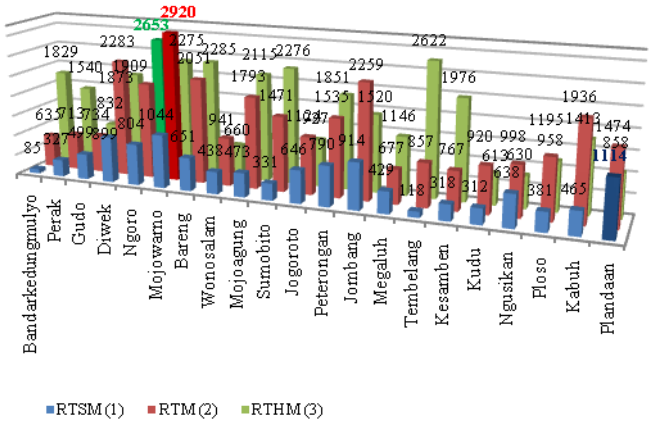
Berikut tahapan yang dilakukan dalam menganalisis data penelitian.

1) Pra-pemprosesan data yang sudah terkumpul (73.720 data) dengan melakukan pengkodingan data pada setiap variabel bertipe kategorik sesuai dengan pengkategorian yang telah ditetapkan. Selain itu juga melakukan *cleaning* data terhadap data-data pengamatan tidak diisi secara lengkap sehingga menyebabkan banyak informasi yang kurang dari unit pengamatan tersebut.
2) Menyajikan statistik deskriptif data untuk memberikan gambaran awal tentang karakteristik data penelitian dan variabel yang diteliti.
3) Membagi sejumlah data hasil tahapan pra-pemrosesan data (diperoleh sebanyak 43.544 data) menjadi data *learning* dan data *testing* dengan kombinasi proporsi tertentu (mengacu dari penelitian-penelitian sebelumnya) yaitu 95 persen : 5 persen, 90 persen : 10 persen, 85 persen : 15 persen, 80 persen : 20 persen, 75 persen : 25 persen dan 70 persen : 30 persen. Masing-masing kombinasi proporsi tersebut diolah untuk mencobakan alternatif metode pemilahan *test sample estimate*. Sehingga diperoleh suatu nilai ketepatan klasifikasi dan banyak simpul terminal yang terbentuk dari masing-masing kombinasi proporsi data *learning-testing* tersebut.
4) Mencoba alternatif lain metode pemilahan yaitu *cross validation V-fold estimate*, dimana dalam penelitian ini nilai *V* yang digunakan adalah 10. Sejumlah data yang sudah fix hasil dari *cleaning* dan pra prosesing data (43.544 data) diolah dengan metode pemilahan *cross validation 10-fold* dengan indeks Gini dan indeks Twoing. Sehingga diperoleh suatu nilai ketepatan klasifikasi dan banyak simpul terminal yang terbentuk.
5) Membandingkan nilai ketepatan klasifikasi data *testing* hasil langkah 5) dan 6). Metode pemilahan yang menghasilkan nilai ketepatan klasifikasi data *testing* terbesar dengan jumlah simpul terminal relatif sederhana adalah yang nantinya dipilih dalam pembuatan pohon klasifikasi maksimal.
6) Melakukan analisis pembentukan pohon klasifikasi maksimal.
7) Menentukan ukuran pohon klasifikasi yang layak dengan melihat besarnya nilai kompleksitas pohon klasifikasi yang terbentuk dan nilai *resubtitution relative cost*. Jika nilai kompleksitas 0,000 dan nilai *resubtitution relative cost* yang kecil (menunjukkan struktur data dari pohon klasifikasi maksimal kompleks) maka perlu dilakukan pemangkasan pohon klasifikasi maksimal (*pruning*) sehingga diperoleh suatu pohon klasifikasi optimal.
8) Melakukan analisis pohon klasifikasi optimal yang terbentuk.
9) Mendapatkan karakteristik kelas simpul terminal-simpul terminal yang dihasilkan dari penelusuran pohon klasifikasi optimal.
10) Menghitung nilai 1-APER yang dihasilkan oleh data *learning* dan data *testing* dari pohon klasifikasi optimal untuk melihat kebaikan dan keakuratan pohon klasifikasi optimal tersebut.

IV. ANALISIS DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini memberikan informasi bahwa banyaknya setiap kelas rumah tangga miskin di masing-masing kecamatan tidak sama. Diantara dua puluh satu kecamatan di kabupaten Jombang, ada sebelas kecamatan yang memiliki jumlah rumah tangga miskin paling banyak termasuk ke dalam kelas RTHM dibandingkan kelas RTM dan RTSM. Hal ini ditunjukkan dengan gambar diagram hijau yang tingginya melebihi diagram batang merah dan biru dalam satu kecamatan. Sebelas kecamatan tersebut meliputi Kecamatan Bandarkedungmulyo, Perak, Gudo, Ngoro, Bareng, Mojoagung, Sumobito, Peterongan, Megaluh, Tembelang, dan Kesamben. Sementarasepuluh kecamatan lainnya memiliki jumlah rumah tangga miskin paling banyak termasuk ke dalam kelas RTM dibandingkan kelas RTHM dan RTSM,ditunjukkan dengan gambar diagram merah yang

tingginya melebihi diagram batang hijau dan biru dalam satu kecamatan. Sepuluh kecamatan tersebut meliputi Kecamatan Diwek, Mojowarno, Wonosalam, Jogoroto, Jombang, Kudu, Ngusikan, Ploso, Kabuh, dan Plandaan. Rumah tangga miskin dengan kelas RTSM paling banyak ada di kecamatan Plandaan (biru tua), kelas RTM paling banyak ada di kecamatan Mojowarno (merah tua), dan kelas RTHM paling banyak ada di kecamatan Ngoro (hijau tua).



Gambar 1 Banyaknya rumah tangga miskin pada setiap kelas per kecamatan di Kabupaten Jombang.

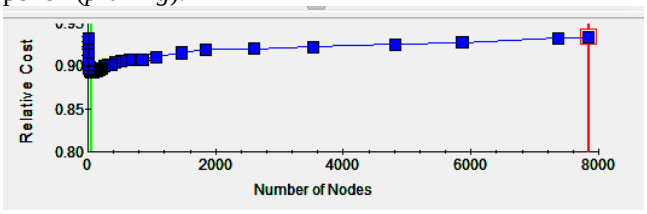
Secara keseluruhan, data yang diperoleh dari survei verifikasi rumah tangga miskin di Kabupaten Jombang tahun 2010 pada kenyataannyamenunjukkan banyak ditemukan data-data *missing* (tidak lengkap) dan data yang *outlier*. Sebelum dilakukan analisis klasifikasi, terlebih dahulu dilakukan pra pemrosesan data dengan cara melakukan filter dan *sorting* data pada masing-masing variabel dan unit pengamatan. Hasil pra-pemrosesan data diperoleh data pengamatan yang akan dianalisis klasifikasi ada sebanyak 43.544 unit rumah tangga miskin.

Tabel 3.
Perbandingan Ketepatan Klasifikasi dari Metode yang Mungkin

	Learning	*Testing*	% ketepatan klasifikasi		Jml. simpul (*Node*)
			Learning	*Testing*	
***Test Sample* dengan Indeks Gini**	95%	5%	76,9	36,2	7.115
	90%	10%	40,5	33,2	23
	85%	15%	34,5	28,3	6
	80%	20%	38,2	33,4	7
	75%	25%	37,4	36,1	3
	70%	30%	35,9	34,1	8
***Test Sample* dengan Indeks Twoing**	95%	5%	67,4	36,5	3.781
	90%	10%	68,1	37,1	3.686
	85%	15%	37,7	29,6	6
	80%	20%	38,2	33,4	7
	75%	25%	40,8	33	80
	70%	30%	43,5	37,2	119
Cross Validation* (CV) 10-*fold	Indeks Gini		43,4	39,7	134
	Indeks Twoing		41	39,7	53

Selanjutnya dilakukan analisis klasifikasi status kemiskinan rumah tangga miskin di Kabupaten Jombang tahun 2010 dengan pohon klasifikasi (CART). Sebelumnya perlu dicari dulu metode terbaik untuk pembentukan pohon klasifikasi yaitu yang menghasilkan ketepatan klasifikasi data *testing* tertinggi. Pada penelitian ini, *test sample* dicoba dengan enam kombinasi data *learning* dan *testing* sebagaimana ditampilkan Tabel 3. Masing-masing dicoba menggunakan fungsi keheterogenan indeks Gini dan indeks Twoing. Persentase ketepatan klasifikasi data *testing* dengan metode pemilah CV menghasilkan nilai yang lebih besar daripada nilai-nilai ketepatan klasifikasi dengan metode t*est sample* pada semua kombinasi data. Ketepatan klasifikasi CV dengan indeks Gini maupun indeks Twoing bernilai sama yaitu 39,7 persen namun jumlah simpul terminal jika digunakan indeks Twoing lebih sedikit (ada 53 simpul) daripada indeks Gini (ada 134 simpul). Berdasarkan konsep parsimony, dalam penelitian ini dipilih metode CV dengan indeks Twoing untuk pembentukan pohon klasifikasi rumah tangga miskin di Kabupaten Jombang tahun 2010.

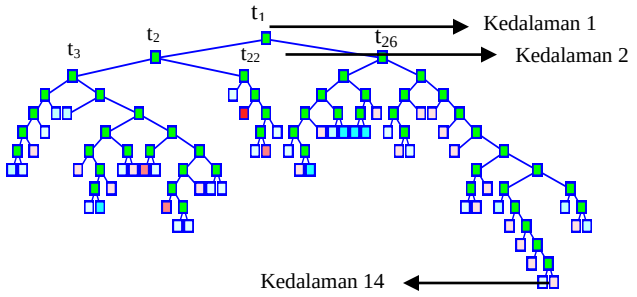
Pohon klasifikasi maksimal yang terbentuk sangat besar dan kompleks dengan simpul terminal sebanyak 7.842 simpul. Nilai kompleksitas yang dihasilkan yaitu 0,000 dengan nilai penduga pengganti (*resubtitution relative cost*) sebesar 0,296 dan biaya kesalahan sebesar 0,934 ± 0,004 (antara 0,930 sampai 0,938). Pada umumnya, pohon klasifikasi maksimal yang telah terbentuk dengan simpul terminal paling banyak bisa jadi akan mengakibatkan terjadinya kasus *underfit* ataupun *overfit* jika ukuran pohon tersebut belum *fit* atau layak.Untuk menghindari itu maka perlu ditentukan pohon klasifikasi dengan ukuran layak melalui proses pemangkasan pohon (*pruning*).



Gambar 2. Plot *relative cost* dengan jumlah simpul tertentu.

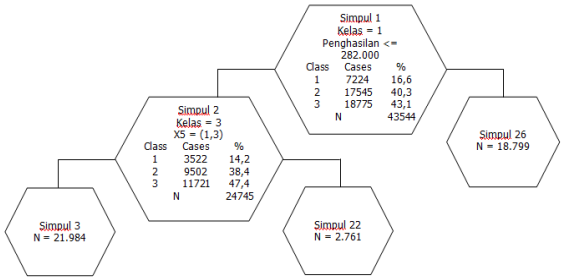
Pemangkasan pohon dilakukan dengan metode *cross validation 10-fold estimate*. Gambar 2 menampilkan nilai*relative cost* pohon klasifikasi maksimal dengan 7.842 simpul termnal sebesar 0,934 (garis merah),sedangkan nilai *relative cost* pohon klasifikasi yang dianggap optimal dengan 53 simpul terminal sebesar 0,894 (garis hijau).Nilai kompleksitas pohon klasifikasi optimal dari hasil pengolahan data diperoleh sebesar 0,000364 dan biaya kesalahan sebesar 0.894 ± 0.004 atau antara 0,890 sampai 0,898.Pemangkasan pohon klasifikasi maksimal menghasilkan pohon klasifikasi optimal yang dianggap sebagai ukuran pohon yang layak untuk klasifikasi status kemiskinan rumah tangga miskin di kabupaten Jombang tahun 2010. dengan jumlah simpul terminal sebanyak 53 simpul dan kedalaman pohon 14. Topologi pohon klasifkikasi optimal tersebut ditampilkan Gambar 3.

Variabel terpenting dalam pemilahan pohon klasifikasi optimal yang mengklasifikasikan rumah tangga miskin di Kabupaten Jombang tahun 2010 berturut-berturut yaitu X_{17} (penghasilan rata-rata per bulan) dengan skor 100, X_{18} (aset yang dimiliki dalam satuan Rupiah) dengan skor 80,29, X_3 (luas lantai) dengan skor 61,73, dan X_8 (tempat pembuangan akhir tinja) dengan skor 60,18. Variabel prediktor yang lain mendapatkan skor di bawah skor keempat variabel tersebut.



Gambar 3. Topologi pohon klasifikasi optimal.

Berikut diberikan visualisasi potongan struktur pohon klasifikasi optimal dari Gambar 3. untukmemberikan contoh penjelasan pemilahan simpul, mulai dari pemilahan simpul utama (simpul 1 atau t_1) menjadi simpul 2 (t_2) dan simpul 26 (t_{26}) dan pemilahan simpul 2 menjadi simpul 3(t_3) dan simpul 22 (t_{22}). Dengan demikian diharapkan agar interpretasi struktur pohon klasifikasi yang terbentuk bisa lebih mudah dipahami secara nyata atau jelas.



Gambar 4. Potongan struktur pohon klasifikasi optimal untuk visualisasi interpretasi.

Variabel X_{17} (penghasilan) memilah simpul utama (simpul 1) menjadi simpul kiri dan simpul kanan dengan ketentuan rumah tangga miskin yang berpenghasilan ≤ Rp 282.000,- akan dipilah menjadi simpul kiri (simpul 2) sedangkan rumah tangga miskin yang berpenghasilan >Rp 282.000,- akan dipilah menjadi simpul kanan (simpul 26). Diperoleh hasil bahwa ada sebanyak 24.745 rumah tangga miskinmemiliki penghasilan ≤ Rp 282.000,- yang menjadi anggota simpul kiri (simpul 2) dan sisanya yaitu 18.799 rumah tangga miskinberpenghasilan >Rp 282.000,- menjadi anggota simpul kanan (simpul 26). Simpul 2 yang beranggo-takan 24.745 rumah tangga miskin dengan penghasilan ≤ Rp 282.000,-selanjutnya dipilah menjadi simpul baru kiri dan kanan menurut jenis dinding terluas (X_5). Jika jenis dinding terluas rumah tangga miskin yaitu tembok ataupun bambu (kategori 1 ataupun 3) maka rumah tangga miskin tersebut akan dipilah menjadi simpul kiri baru (simpul 3). Namun jika jenis dinding terluasnya adalah kayu (kategori 2) maka akan dipilah menjadi anggota simpul kanan baru (simpul 22). Diantara 24.745 rumah tangga miskin anggota simpul 2, diperoleh hasil ada sebanyak 21.984 rumah tangga miskin yang menjadi anggota simpul 3 dengan karakteristik penghasilan ≤ Rp 282.000,- dan jenis dinding terluas yaitu tembok atau bambu (kategori 1 atau 3). Sisanya ada 2.761 rumah tangga miskin yang menjadi anggota simpul 22 dengan karakteristik penghasilan ≤ Rp 282.000,- dan jenis dinding terluas yaitu kayu (kategori 2).

Hasil perhitungan keakurasian klasifikasi dari Tabel 4 diperoleh senilai 40,986 persen. Artinya bahwa pohon klasifikasi optimal mampu mengklasifikasikan suatu rumah tangga miskin ke dalam kelas-kelas rumah tangga miskin (RTSM, RTM atau RTHM) dengan tepat sebesar 40,986 persen. Sementara keakuratan klasifikasi data *testing* diperoleh sebesar 39,654 persen. Artinya bahwa pohon klasifikasi optimal memiliki keakuratan hasil prediksi suatu rumah tangga miskin termasuk ke dalam salah satu kelas variabel respon sebesar 39,654 persen.

Tabel 4.
Keakuratan Klasifikasi Data *Learning*oleh Pohon Klasifikasi Optimal

Kelas Aktual	Kelas Prediksi		
	RTSM	RTM	RTHM
RTSM	3.597	1.205	2.422
RTM	6.250	4.157	7.138
RTHM	4.995	3.687	10.093

Tabel 5.
Keakuratan Klasifikasi Data *Testing*oleh Pohon Klasifikasi Optimal

Kelas Aktual	Kelas Prediksi		
	RTSM	RTM	RTHM
RTSM	3.272	1.346	2.606
RTM	6.102	3.848	7.595
RTHM	5.111	3.517	10.147

V. KESIMPULAN

Diantara 73.720 rumah tangga miskin di Kabupaten Jombang tahun 2010, sekitar 15,8 persen termasuk kelas RTSM, sebanyak 39,6 persen termasuk RTM, dan paling banyak yaitu sekitar 44,6 persen termasuk kelas RTHM. Variabel terpenting atau paling dominan berpengaruh dalam menentukan status kemiskinan suatu rumah tangga miskin di Kabupaten Jombang tahun 2010 yaitu penghasilan rata-rata per bulan (Rp). Data sampel *learning* secara keseluruhan tepat diklasifikasikan oleh pohon klasifikasi yang dihasilkan sebesar 40,986 persen dan akurasi prediksi data *testing* sebesar 39,654 persen.

Untuk meningkatkan tingkat akurasi klasifikasi pada penelitian berikutnya, data yang akan dianalisis perlu disiapkan dengan benar pada pra pemrosesan data serta bisa dicobakan metode *ensamble* CART sebagai metode alternatif.

DAFTAR PUSTAKA

[1] Suyanto. 2010. http://jombangkab.go.id/egov/layanan/ berita.asp? menu =detailberita&no =2200. diakses pada tanggal 25 Oktober 2013.

[2] [BPS PPLS] Badan Pusat Statistik. 2011. *Analisis Data Kemiskinan Berdasarkan Data Pendataan Program Perlindungan Sosial* 2011.[http://www.bps.go .id], diakses pada tanggal 29 Desember 2013.

[3] Breiman L, Friedman J, Olshen R, dan Stone C.1993. *Classification and Regression Trees*. Chapman Hall: New York–London.

[4] Muttaqin, J. 2013. *Metode Ensamble CART untuk Perbaikan Klasifikasi Kemiskinan di Kabupaten Jombang*. ITS Press: Surabaya.

[5] Lewis, R J. 2000. *An Introduction to Classification And Regression Trees (CART) Analysis*, Departement of Emergency Medicine Harbor-UCLA Medical Center, Torrance, California.

[6] Johnson, R.A. dan Wichern, D.W. 1992. *Applied Multivariate Statistical Analysis*, Prentice Hall: New Jersey.

[7] [BPS SUSENAS] Badan Pusat Statistik.2008. *Survei Sosial Ekonomi Nasional (SUSENAS) Tahun 2008*. BPS: Jakarta.

[8] *World Bank Institute*. 2002. *Dasar-dasar Analisis Kemiskinan*. Jakarta.