IJSA (2024), Vol. 8, No.1, 1-13
DOI: https://doi.org/10.29244/ijsa.v8i1p1-13

Indonesian Journal of Statistics and Its Applications
Available at https://journal-stats.ipb.ac.id/index.php/ijsa
e-ISSN: 2599-0802

Comparison of Chi-Square Automatic Interaction Detector (CHAID) and Random Forest Methods in the Classification of Household Poverty Status in Central Java*

Perbandingan Metode Chi-Square Automatic Interaction Detector (CHAID) dan Random Forest dalam Klasifikasi Status Kemiskinan Rumah Tangga di Jawa Tengah

Fatkhul Izzati[1], Mohammad Masjkur[2‡], Farit Mochamad Afendi[3]

[1,2,3]Department of Statistics, IPB University, Indonesia
‡corresponding author: masjkur@apps.ipb.ac.id

Abstract

Central Java was in second position as the province with the highest number of poor people in Indonesia in March 2020. Poverty alleviation efforts have been carried out, but many are still not yet on target. The purpose of this study was to model the classification of household poverty status in Central Java using CHAID and random forest methods and compare the two methods. The data used in this study is data from the 2020 National Socioeconomic Survey (SUSENAS) conducted by the Central Bureau of Statistics (BPS) for Central Java. The number of poor households is much less than non-poor households. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) was performed to handle unbalanced data. The random forest method produced better classification performance than the CHAID method with accuracy, sensitivity, specificity, and AUC of 93,95%, 98,43%, 89,92%, and 0,9417, respectively. The important variables that build the random forest model are the floor area of the house, the age of the head of the household, cooking fuel, the place for the final disposal of feces, and ownership of the place to defecate.

Keywords: chaid, poverty, random forest, smote.

* Received: Nov 2022; Reviewed: Feb 2024; Published: Jun 2024

1. Pendahuluan

Jawa Tengah menduduki posisi kedua sebagai provinsi dengan jumlah penduduk miskin terbanyak di Indonesia. Menurut data Badan Pusat Statistik (BPS), pada bulan Maret 2020 jumlah penduduk miskin di Jawa Tengah mencapai 3,98 juta orang (11,41 persen) (BPS, 2020). Angka tersebut meningkat sebesar 301,5 ribu orang dibandingkan dengan jumlah penduduk miskin pada September 2019 yang berjumlah 3,68 juta orang (10,58 persen). Sebelumnya, Pemerintah Provinsi Jawa Tengah telah menargetkan penurunan angka kemiskinan di Jawa Tengah menjadi 9,81 persen di tahun 2020. Berbagai upaya untuk menanggulangi masalah kemiskinan sudah dilakukan oleh pemerintah, seperti pemberian Bantuan Langsung Tunai (BLT) maupun bentuk bantuan lainnya. Namun, masih banyak bantuan yang diberikan tidak tepat sasaran, seperti BLT dan Bantuan Sosial (Bansos). Untuk membantu pemerintah dalam mengatasi masalah tersebut, perlu adanya semacam model yang dapat mengklasifikasikan status kemiskinan rumah tangga miskin dan tidak miskin.

Klasifikasi merupakan analisis statistika yang dapat digunakan untuk mengelompokkan suatu objek ke dalam kelompok tertentu yang sesuai dengan nilai peubah-peubahnya. Kristiani mengklasifikasikan kelompok rumah tangga di Kabupaten Blora menggunakan Multivariate Adaptive Regression Spline (MARS) dan Fuzzy K-Nearest Neighbor (FK-NN) (Kristiani et al., 2015), sedangkan Nuzula mengklasifikasikan status kemiskinan rumah tangga di Kabupaten Wonosobo menggunakan metode Support Vector Machines (SVM) dan Classification and Regression Trees (CART) (Nuzula et al., 2020). Metode lain yang dapat digunakan dalam pengklasifikasian diantaranya metode Chi-square Automatic Interaction Detection (CHAID) dan random forest.

Metode CHAID merupakan suatu metode eksplorasi yang dapat digunakan untuk mengetahui hubungan antara peubah respon dengan peubah penjelas, selain itu juga secara otomatis dapat mendeteksi adanya interaksi antar peubah penjelas. Kelebihan dari metode ini adalah hasilnya berupa dendogram sederhana yang dapat memberikan informasi mengenai hubungan terstruktur antara peubah respon dengan peubah penjelas yang mudah dimengerti.

Random forest merupakan metode klasifikasi yang berasal dari pengembangan metode Classification and Regression Tree (CART) yang menggabungkan antara metode bootstrap aggregating bagging dan random feature selection (Breiman, 2001). Metode pohon gabungan memiliki kemampuan pendugaan dan akurasi yang lebih baik dibandingkan dengan pohon tunggal (Sartono & Syafitri, 2010).

Akurasi yang diperoleh pada penelitian-penelitian sebelumnya menggunakan metode CHAID dan random forest sudah cukup baik. Yanthy dalam Penentuan Karakteristik Kelancaran Pembayaran Kartu Kredit Menggunakan Metode CHAID memperoleh akurasi 80,8% (Yanthy, 2013). Oktavia dalam Faktor-Faktor yang Berpengaruh dalam Mendapatkan Pekerjaan Bagi Lulusan Statistika IPB dengan Menggunakan Metode CHAID (Studi Kasus : Alumni Departemen Statistika IPB Angkatan 48-50) memperoleh akurasi 79,3% (Oktavia, 2018). Sulviana dalam Penggunaan Metode CHAID pada Segmentasi Tren Penjualan Berbagai Jenis Minuman Ringan di Indonesia memperoleh ketepatan klasifikasi sebesar 71,4% (Sulviana, 2018). Utami dalam Penerapan Metode Random Forest dalam Menentukan Status Istitaah Kesehatan Jemaah Haji memperoleh akurasi sebesar 89,8% (Utami,

2019). Hidayat memperoleh nilai akurasi random forest sebesar 97,26% dalam Sistem Penunjang Keputusan untuk Identifikasi Kelelawar Menggunakan Random Forest dan C5.0 (Hidayat, 2016). Nugraha dalam Pendeteksian Lalu Lintas Botnet Berbasis Jaringan dengan K-Nearest Neighbor dan Random Forest memperoleh nilai akurasi random forest sebesar 98,41% (Nugraha, 2017).

Melalui metode CHAID dan random forest akan dilakukan klasifikasi status kemiskinan rumah tangga di Jawa Tengah, selanjutnya akan dibandingkan metode mana yang lebih baik dalam melakukan pemodelan

2. Metodologi

2.1 Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang berasal dari Survei Sosial Ekonomi Nasional (SUSENAS) 2020 yang dilakukan oleh Badan Pusat Statistik (BPS) untuk wilayah Provinsi Jawa Tengah.

Peubah yang diamati dalam penelitian ini terdiri atas satu peubah respon (Y) dan 12 peubah penjelas (X) yang merupakan peubah signifikan dalam pengklasifikasian status kemiskinan rumah tangga di Jawa Barat (Nurpadilah, 2019). Peubah-peubah tersebut dapat dilihat pada Tabel 1.

Tabel 1: Peubah yang digunakan dalam penelitian

Kode	Peubah	Keterangan
Y	Status kemiskinan rumah tangga	1: Rumah tangga miskin 0: Rumah tangga tidak miskin
X1	Kepemilikan sepeda motor	1: Ya 0: Tidak
X2	Luas lantai rumah (m^2)	3 sampai 800 m^2
X3	Kepemilikan tempat buang air besar	1: Milik sendiri 2: Milik bersama 3: Umum 4: Tidak ada
X4	Tempat pembuangan akhir tinja	1: Tangki septik 2: Kolam/sawah/sungai/danau/laut 3: Lubang tanah 4: Tidak ada fasilitas 5: Lainnya
X5	Umur kepala rumah tangga	15 sampai 97 tahun
X6	Pernah menerima bantuan pangan	1: Ya 0: Tidak
X7	Kepemilikan lemari es	1: Ya 0: Tidak
X8	Bahan bakar memasak	1: Elpiji 2: Kayu bakar 3: Gas kota 4: Tidak memasak 5: Lainnya
X9	Banyaknya anggota rumah tangga	1: Anggota rumah tangga ≤ 4 2: 5 ≤ Anggota rumah tangga ≤ 7

Kode	Peubah	Keterangan
		3: Anggota rumah tangga > 7
X10	Jenjang pendidikan terakhir kepala rumah tangga	1: Tidak pernah bersekolah 2: SD/MI/sederajat 3: SMP/MTS/sederajat 4: SMA/MA/sederajat 5: Perguruan tinggi
X11	Sumber utama air minum	1: Air kemasan atau isi ulang 2: Sumur 3: Mata air 4: Leding 5: Lainnya
X12	Sumber utama air untuk memasak dan mandi	1: Air kemasan atau isi ulang 2: Sumur 3: Mata air 4: Leding 5: Lainnya

2.2 Metode Penelitian

Tahapan analisis data dalam penelitian sebagai berikut:

1) Mengklasifikasikan status kemiskinan rumah tangga sesuai dengan definisi BPS. Rumah tangga miskin adalah rumah tangga yang memiliki rata-rata pengeluaran per kapita per bulan di bawah garis kemiskinan. Garis kemiskinan yang digunakan pada penelitian ini adalah Garis Kemiskinan Provinsi Jawa Tengah pada tahun 2020.
2) Melakukan eksplorasi data dan analisis statistika deskriptif untuk mengetahui gambaran umum peubah-peubah yang akan dianalisis.
3) Menangani data tidak seimbang dengan metode SMOTE.
4) Melakukan pemodelan dengan metode CHAID dengan tahapan:
    a. Mengubah peubah numerik menjadi peubah kategorik.
    b. Melakukan pemodelan dengan metode CHAID.
5) Melihat performa model klasifikasi CHAID menggunakan validasi silang 10-fold dengan tahapan:
    a. Membagi data menjadi 10 bagian secara acak dengan jumlah amatan yang relatif sama di setiap bagian dan proporsi kelas yang relatif sama dengan data latih utama.
    b. Satu bagian digunakan sebagai data validasi dan sembilan bagian lainnya digunakan sebagai data latih. Tahap ini dilakukan secara bergantian sehingga terdapat 10 data latih dan 10 data validasi.
6) Melakukan pemodelan dengan metode random forest.
7) Melihat performa model klasifikasi random forest menggunakan validasi silang 10-fold seperti pada tahap 5.
8) Mengevaluasi model yang dihasilkan pada tahapan 4 dan 7 dengan membandingkan nilai akurasi, sensitivitas, spesifisitas, AUC (Area under the ROC Curve) dan peubah penting yang dihasilkan dari model.

3. Hasil dan Pembahasan

3.1 Deskripsi Data

Data yang digunakan dalam penelitian ini berjumlah 29190 amatan rumah tangga yang terdiri atas 2709 rumah tangga miskin dan 26481 rumah tangga tidak miskin. Jumlah rumah tangga miskin jauh lebih sedikit dibandingkan dengan rumah tangga tidak miskin. Gambaran umum kategori status kemiskinan rumah tangga di Jawa Tengah dapat dilihat pada Tabel 2.

Tabel 2: Gambaran umum kategori status kemiskinan rumah tangga di Provinsi Jawa Tengah

Status Kemiskinan	Frekuensi	Persentase
Miskin	2709	9,28%
Tidak Miskin	26481	90,72%

Semakin banyak anggota rumah tangga maka rumah tangga tersebut memiliki kecenderungan yang lebih besar untuk mengalami kemiskinan dibandingkan dengan rumah tangga yang memiliki anggota rumah tangga lebih sedikit sebagaimana ditunjukkan pada Tabel 3.

Tabel 3: Persentase status kemiskinan berdasarkan banyaknya anggota rumah tangga

Status Kemiskinan	ART≤4	5≤ART≤7	ART>7
Miskin	7%	16%	25%
Tidak miskin	93%	84%	75%

Kepala rumah tangga yang tidak pernah bersekolah memiliki kecenderungan yang lebih tinggi untuk mengalami kemiskinan dibandingkan dengan kepala rumah tangga yang pernah bersekolah. Semakin tinggi jenjang pendidikan yang pernah ditempuh oleh kepala rumah tangga maka semakin kecil kecenderungan rumah tangga tersebut mengalami kemiskinan.

Tabel 4: Persentase status kemiskinan berdasarkan jenjang pendidikan terakhir kepala rumah tangga

Status Kemiskinan	Tidak sekolah	SD	SMP	SMA	PT
Miskin	17%	12%	8%	5%	1%
Tidak miskin	83%	88%	92%	95%	99%

3.2 Penanganan Data Tidak Seimbang

Menurut Chawla kondisi data yang tidak seimbang akan mengakibatkan pengklasifikasian cenderung diliputi oleh kelas mayor dan mengabaikan kelas minor (Chawla et al., 2002). Oleh karena itu, perlu dilakukan penanganan data tidak seimbang. Metode yang digunakan dalam penelitian ini adalah Synthetic Minority Oversampling Technique (SMOTE). Metode SMOTE memperbanyak amatan dari kelas minor agar jumlahnya setara dengan jumlah amatan yang ada di kelas mayor dengan membuat data baru berdasarkan k-tetangga terdekat (k-nearest neighbor). Nilai k yang digunakan dalam penelitian ini sebesar k = 5 dengan persentase

oversampling sebesar 900% dan persentase undersampling sebesar 100%. Hasil dari proses SMOTE disajikan pada Tabel 5.

Berdasarkan Tabel 5 terlihat bahwa data hasil SMOTE menjadi lebih seimbang sehingga siap digunakan untuk tahap klasifikasi.

Tabel 5: Hasil SMOTE pada data latih

Keterangan	Sebelum SMOTE		Setelah SMOTE	
	Miskin	Tidak Miskin	Miskin	Tidak Miskin
Data latih utama	2709	26481	27090	24381

3.3 Pemodelan dengan Metode CHAID

Sebelum dilakukan pemodelan dengan metode CHAID, peubah numerik harus diubah terlebih dahulu ke dalam peubah kategorik. Terdapat dua peubah numerik yaitu peubah X2 (luas lantai rumah (m^2)) dan peubah X5 (umur kepala rumah tangga). Pengkategorian peubah yang dilakukan dalam penelitian ini berdasarkan referensi dari situs web resmi dari BPS, pengkategorian tersebut disajikan pada Tabel 6 dan Tabel 7.

Tabel 6: Pengkategorian peubah X2 (luas lantai rumah (m^2))

Luas lantai (m^2)	Kategori
≤19	1
20−49	2
50−99	3
100−149	4
≥150	5

Tabel 7: Pengkategorian peubah X5 (umur kepala rumah tangga)

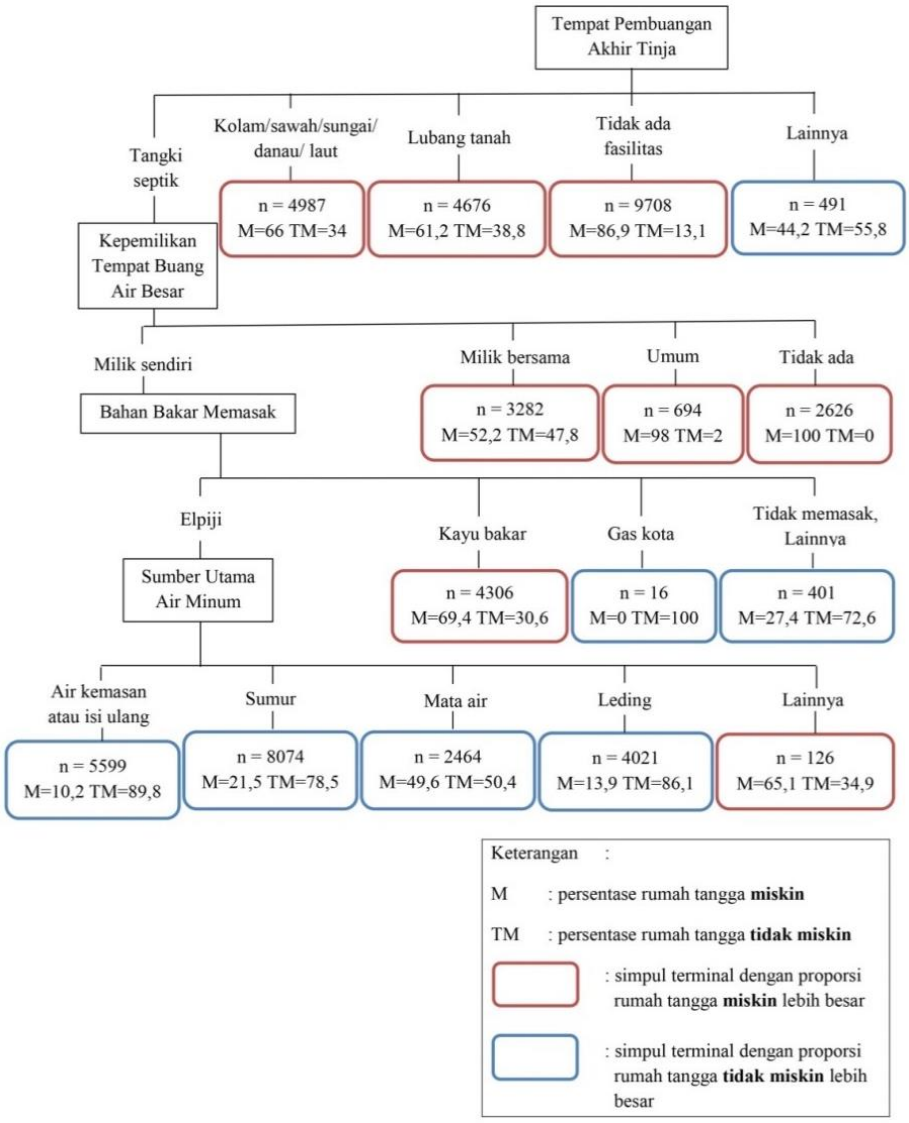
Umur	Kategori
≤24	1
25−44	2
45−59	3
≥60	4

Setelah semua peubah sudah berupa peubah kategorik, tahap klasifikasi dengan metode CHAID dapat dilakukan.

Pembentukan model klasifikasi dilakukan dengan menggunakan data latih utama, yaitu data keseluruhan yang sudah dilakukan penanganan data tidak seimbang menggunakan metode SMOTE. Hasil yang diperoleh berupa pohon klasifikasi. Pembentukan pohon klasifikasi CHAID secara umum meliputi tiga hal, yaitu tahap penggabungan (merging), tahap pemisahan (splitting), dan tahap penghentian (stopping) (Kass, 1980). Kriteria uji statistik yang digunakan adalah kriteria uji statistik khi-kuadrat dengan taraf nyata 5%. Kategori peubah dengan nilai uji khi-kuadrat terbesar akan dijadikan simpul utama yang merupakan peubah paling signifikan dalam analisis CHAID. Pohon klasifikasi dibentuk dengan batasan minsplit sebesar 10000 agar pohon klasifikasi yang terbentuk tidak terlalu besar dan lebih mudah untuk diinterpretasikan. Minsplit adalah opsi untuk menentukan berapa ukuran node minimal yang diperbolehkan untuk melakukan pemisahan (splitting). Minsplit = 10000 artinya, jika ada simpul yang amatannya kurang dari 10000 maka proses splitting akan

dihentikan dan algoritmanya berhenti.

Pohon klasifikasi CHAID yang terbentuk dapat dilihat pada Gambar 1. Peubah yang paling signifikan adalah peubah tempat pembuangan akhir tinja, diikuti dengan peubah kepemilikan tempat buang air besar, bahan bakar memasak dan sumber utama air minum.



Gambar 1: Pohon klasifikasi CHAID

Kebaikan model klasifikasi metode CHAID dapat dilihat dengan validasi silang q-fold (q-fold cross validation). Menurut James banyaknya q yang digunakan biasanya $q = 5$ dan $q = 10$ (James et al., 2013). Pada penelitian ini dilakukan validasi silang 10-fold. Performa hasil klasifikasi melalui validasi silang 10-fold dapat dilihat pada Tabel 8.

Spesifisitas merupakan kemampuan model dalam memprediksi rumah tangga miskin, sedangkan sensitivitas adalah kemampuan model dalam memprediksi rumah tangga tidak miskin. Akurasi menggambarkan tingkat ketepatan klasifikasi secara keseluruhan. Nilai AUC didapat dari luas daerah di bawah kurva ROC. Kurva ROC menunjukkan kinerja model klasifikasi melalui grafik dua dimensi, sumbu x

menunjukkan nilai peluang salah positif (1-spesifisitas) dan sumbu y menunjukkan nilai prediksi benar positif (sensitivitas). Nilai akurasi, sensitivitas, spesifisitas, serta AUC yang diperoleh dari masing-masing bagian tidak jauh berbeda. Nilai rataan akurasi sebesar 76,75% artinya model mampu mengklasifikasikan status kemiskinan rumah tangga di Jawa Tengah dengan cukup baik. Berdasarkan nilai rataan AUC yang diperoleh sebesar 0,7643 diklasifikasikan sebagai tingkat akurasi sedang (Gorunescu, 2011).

Tabel 8: Performa model klasifikasi CHAID pada setiap bagian

Fold	Akurasi (%)	Sensitivitas (%)	Spesifisitas (%)	AUC
1	61,43	67,92	55,59	0,6176
2	79,76	79,61	79,88	0,7975
3	77,46	67,64	86,30	0,7679
4	78,08	74,41	81,40	0,7790
5	77,27	68,29	85,35	0,7682
6	77,13	68,01	85,35	0,7668
7	76,78	61,40	90,62	0,7601
8	80,09	74,53	85,09	0,7981
9	78,40	63,47	91,84	0,7765
10	81,14	81,21	81,07	0,8114
Rataan	**76,75**	**70,65**	**82,25**	**0,7643**

3.4 Pemodelan dengan Metode Random Forest

Random forest merupakan metode klasifikasi yang berasal dari pengembangan metode Classification and Regression Tree (CART) yang menggabungkan antara metode bootstrap aggregating (bagging) dan random feature selection (Breiman, 2001). Pohon yang dihasilkan berjumlah banyak sehingga menyerupai hutan (forest). Analisis dilakukan dari kumpulan pohon tersebut.

Data latih diambil dari data latih utama, keseluruhan data yang sudah ditangani menggunakan SMOTE agar seimbang. Pemodelan dilakukan dengan berbagai kombinasi parameter m dan r yang berbeda. Nilai m yang disarankan adalah (Breiman & Cutler, 2003):

$$m = \frac{\left|\sqrt{p}\right|}{2}$$

$$m = \left|\sqrt{p}\right|$$

$$m = 2\left|\sqrt{p}\right|$$

Keterangan:
m : banyak peubah penjelas yang digunakan sebagai peubah pemilah
p : banyak peubah penjelas pada penelitian

Jumlah p yang digunakan sebanyak 12 peubah. Sehingga, didapatkan nilai m yang dicobakan yaitu 2, 3 dan 7.

Nilai $r = 50$ umumnya sudah menghasilkan klasifikasi yang baik (Breiman, 1996). Sutton menyarankan untuk menggunakan nilai $r \geq 100$ untuk menghasilkan misklasifikasi yang relatif stabil (Sutton, 2005). Pada penelitian ini, nilai r yang dicobakan yaitu 25, 50, 100, dan 200.

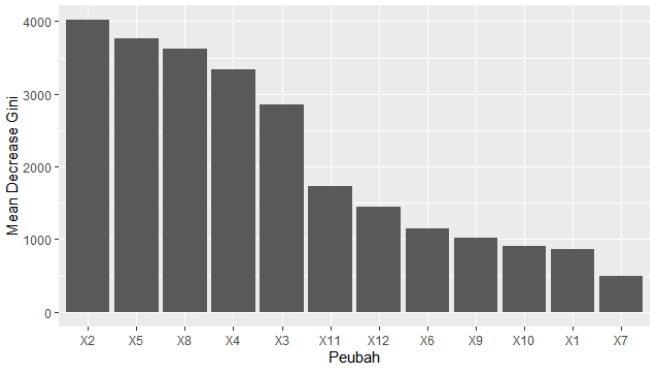
Nilai akurasi, sensitivitas, spesifisitas, dan AUC yang dihasilkan dari masing-masing kombinasi m dan r disajikan dalam Tabel 9.

Performa model klasifikasi random forest terbaik didapat dari kombinasi $m = 7$ dan $r = 200$ dengan nilai akurasi, sensitivitas, spesifisitas dan AUC yang terbesar dibandingkan dengan kombinasi m dan r lainnya.

Tabel 9: Performa model klasifikasi random forest dalam berbagai kombinasi nilai m dan r

Ket		m					
		2		3		7	
r	25	Akurasi	83,87%	Akurasi	88,40%	Akurasi	92,47%
		Sensitivitas	87,26%	Sensitivitas	92,07%	Sensitivitas	92,73%
		Spesifisitas	50,72%	Spesifisitas	52,49%	Spesifisitas	89,92%
		AUC	0,6899	AUC	0,7228	AUC	0,9132
	50	Akurasi	83,78%	Akurasi	88,11%	Akurasi	92,85%
		Sensitivitas	87,21%	Sensitivitas	91,57%	Sensitivitas	93,03%
		Spesifisitas	50,28%	Spesifisitas	54,23%	Spesifisitas	91,10%
		AUC	0,6874	AUC	0,7290	AUC	0,9207
	100	Akurasi	83,98%	Akurasi	88,50%	Akurasi	92,83%
		Sensitivitas	87,36%	Sensitivitas	91,96%	Sensitivitas	92,93%
		Spesifisitas	50,98%	Spesifisitas	54,67%	Spesifisitas	91,81%
		AUC	0,6917	AUC	0,7331	AUC	0,9237
	200	Akurasi	83,49%	Akurasi	88,53%	Akurasi	93,03%
		Sensitivitas	86,76%	Sensitivitas	91,93%	Sensitivitas	93,07%
		Spesifisitas	51,50%	Spesifisitas	55,22%	Spesifisitas	92,62%
		AUC	0,6913	AUC	0,7358	AUC	0,9284

Analisis dilanjutkan dengan melihat peubah penting yang membangun model. Peubah penting adalah peubah yang dapat mendiskriminasi kategori respon, yaitu rumah tangga miskin dan tidak miskin. Pada algoritma random forest peubah penting dapat diketahui dengan menghitung nilai mean decrease gini (MDG). Ukuran tersebut digunakan untuk melihat kestabilan tiap peubah bebas dalam random forest. Nilai MDG digunakan sebagai penentu urutan tingkat kepentingan peubah penjelas. Semakin tinggi nilai MDG maka semakin baik (Breiman, 2001), sehingga peubah yang memiliki nilai MDG semakin besar akan menempati tingkat kepentingan yang semakin tinggi dalam menentukan status kemiskinan rumah tangga Jawa Tengah. Urutan tingkat kepentingan peubah penjelas pada random forest terbaik dapat dilihat pada Gambar 2.



Gambar 2: Tingkat kepentingan peubah penjelas pada random forest terbaik

Peubah X2 (luas lantai rumah) menjadi peubah dengan tingkat kepentingan tertinggi, diikuti dengan peubah X5 (umur kepala rumah tangga), X8 (bahan bakar memasak), X4 (tempat pembuangan akhir tinja) dan X3 (kepemilikan tempat buang air besar). Penurunan nilai MDG secara drastis terjadi pada peubah pada peringkat selanjutnya, yaitu X11 (sumber utama air minum), X12 (sumber utama air untuk memasak dan mandi), X6 (pernah menerima bantuan pangan), X9 (banyaknya anggota rumah tangga), X10 (jenjang pendidikan terakhir kepala rumah tangga), X1 (kepemilikan sepeda motor) dan X7 (kepemilikan lemari es).

Kebaikan model klasifikasi random forest terbaik selanjutnya akan dilihat melalui validasi silang 10-fold dengan menghitung nilai akurasi, sensitivitas, spesifisitas dan AUC dari masing-masing bagian. Tabel 10 menunjukkan performa model klasifikasi random forest pada tiap bagian data.

Tabel 10: Performa model random forest dengan $m = 7$ dan $r = 200$ pada setiap bagian

Fold	**Akurasi (%)**	**Sensitivitas (%)**	**Spesifisitas (%)**	**AUC**
1	47,39	100,00	0,00	0,5002
2	99,22	98,36	100,00	0,9918
3	99,38	98,69	100,00	0,9934
4	99,24	98,48	99,93	0,9920
5	99,03	98,03	99,93	0,9898
6	99,01	98,11	99,82	0,9896
7	99,05	98,15	99,85	0,9900
8	98,93	97,79	99,96	0,9887
9	99,16	98,52	99,74	0,9913
10	99,09	98,15	99,93	0,9904
Rataan	93,95	98,43	89,92	0,9417

Nilai akurasi, sensitivitas, spesifisitas dan AUC yang diperoleh dari masing-masing bagian tidak jauh berbeda. Nilai rataan akurasi sebesar 93,95% artinya model mampu mengklasifikasikan status kemiskinan rumah tangga di Jawa Tengah dengan sangat baik. Berdasarkan nilai rataan AUC diklasifikasikan sebagai tingkat akurasi sangat tinggi karena nilai rataan AUC yang diperoleh sebesar 0,9417.

3.5 Pemodelan Model Klasifikasi

Peubah yang paling signifikan dalam model klasifikasi CHAID adalah peubah tempat pembuangan akhir tinja, diikuti dengan peubah kepemilikan tempat buang air besar, bahan bakar memasak dan sumber utama air minum. Sedangkan, peubah yang paling signifikan dalam model klasifikasi random forest adalah peubah luas lantai rumah, diikuti dengan peubah umur kepala rumah tangga, bahan bakar memasak, tempat pembuangan akhir tinja dan kepemilikan tempat buang air besar. Terdapat perbedaan dalam tingkat kepentingan peubah pada kedua model klasifikasi. Peubah yang memiliki pengaruh besar dalam pembangunan kedua model klasifikasi adalah peubah bahan bakar memasak, tempat pembuangan akhir tinja dan kepemilikan tempat buang air besar.

Berdasarkan model yang terbentuk, model CHAID dapat mengetahui bagaimana alur pengambilan keputusan melalui pohon klasifikasi CHAID, dapat dilihat

bagaimana peran masing-masing peubah dan interaksi antar peubahnya dalam mengklasifikasikan status kemiskinan rumah tangga di Jawa Tengah, sedangkan model random forest hanya dapat mengetahui tingkat kepentingan peubah dan hasil klasifikasi yang didapatkan tanpa melihat bagaimana alur pengambilan keputusan tersebut dilakukan. Namun, jika dilihat berdasarkan nilai akurasi, sensitivitas, spesifisitas dan AUC dari masing-masing model, model klasifikasi random forest memiliki performa lebih baik dibandingkan model klasifikasi CHAID. Perbandingan rataan nilai akurasi, sensitivitas, spesifisitas dan AUC dari model CHAID dan random forest dapat dilihat pada Tabel 11.

Tabel 11: Perbandingan performa model CHAID dan random forest

Model Klasifikasi	**Akurasi (%)**	**Sensitivitas (%)**	**Spesifisitas (%)**	**AUC**
CHAID	76,75	70,65	82,25	0,7643
Random Forest	93,95	98,43	89,92	0,9417

Nilai sensitivitas dan spesifisitas model klasifikasi random forest lebih tinggi dibandingkan dengan model klasifikasi CHAID, hal ini berarti model random forest memiliki kemampuan lebih baik dalam memprediksi suatu rumah tangga miskin maupun suatu rumah tangga tidak miskin. Model random forest memperoleh nilai akurasi dan AUC lebih tinggi, nilai tersebut menunjukkan bahwa secara keseluruhan metode random forest memiliki tingkat ketepatan yang lebih tinggi dalam menentukan status kemiskinan rumah tangga di Jawa Tengah dibandingkan model klasifikasi CHAID

4. Simpulan dan Saran

Data yang digunakan dalam penelitian ini merupakan data tidak seimbang sehingga diperlukan penanganan menggunakan metode SMOTE. Analisis klasifikasi dilakukan menggunakan dua metode, yaitu metode CHAID yang merupakan pohon tunggal, serta metode random forest yang merupakan kumpulan dari beberapa pohon.

Peubah yang paling signifikan dalam model klasifikasi CHAID adalah peubah tempat pembuangan akhir tinja, diikuti dengan peubah kepemilikan tempat buang air besar, bahan bakar memasak dan sumber utama air minum. Sedangkan, peubah yang paling signifikan dalam model klasifikasi random forest adalah peubah luas lantai rumah, diikuti dengan peubah umur kepala rumah tangga, bahan bakar memasak, tempat pembuangan akhir tinja dan kepemilikan tempat buang air besar. Peubah yang memiliki pengaruh besar dalam pembangunan kedua model klasifikasi adalah peubah bahan bakar memasak, tempat pembuangan akhir tinja dan kepemilikan tempat buang air besar.

Kelebihan dari model CHAID adalah dapat melihat alur pengambilan keputusan melalui pohon klasifikasi yang terbentuk, sedangkan random forest hanya dapat melihat hasil keputusan akhirnya saja. Namun, bersadarkan nilai akurasi, sensitivitas, spesifisitas dan AUC secara keseluruhan metode random forest memiliki performa yang lebih baik dalam memprediksi status kemiskinan rumah tangga di Jawa Tengah dibandingkan metode CHAID.

Daftar Pustaka

BPS. (2020). Kemiskinan Provinsi Jawa Tengah Maret 2020. Badan Pusat Statistik. https://jateng.bps.go.id/pressrelease/2020/07/15/1225/persentase-penduduk-miskin-maret-2020-naik-menjadi-11-41-persen--dibanding-september-2019--yang-sebesar-10-58-persen.html

Breiman, L. (1996). Bagging Predictors. Machine Learning, 24, 123–140.

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32.

Breiman, L., & Cutler, A. (2003). Manual on setting up, using, and understanding random forest v4.0. https://www.stat.berkele.edu/~breiman/Using_random_forest_v4.0.pdf

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE : synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Gorunescu, F. (2011). Data Mining: Concept, Models and Techniques. Berlin (GER): Springer-Verlag Berlin Heidenberg.

Hidayat, D. S. (2016). Sistem penunjang keputusan untuk identifikasi kelelawar menggunakan random forest dan C5.0. Tesis, Sekolah Pascasarjana, Institut Pertanian Bogor, Bogor.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Application in R. New York (US): Springer.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29(2), 119–127.

Kristiani, Y. P., Safitri, D., & Ispriyanti, D. (2015). Klasifikasi kelompok rumah tangga di Kabupaten Blora menggunakan multivariate adaptive regression spline (MARS) dan fuzzy k-nearest neighbor (FK-NN). Jurnal Gaussian, 4(4), 1077–1085.

Nugraha, H. (2017). Pendeteksian lalu lintas botnet berbasis jaringan dengan k-nearest neighbor dan random forest. Skripsi, Departemen Ilmu Komputer, Institut Pertanian Bogor, Bogor.

Nurpadilah, W. (2019). Metode ensemble pada pohon klasifikasi tunggal untuk klasifikasi status kemiskinan rumah tangga di Provinsi Jawa Barat. Skripsi, Departemen Statistika, Institut Pertanian Bogor, Bogor.

Nuzula, L., Prahutama, A., & Hakim, A. R. (2020). Klasifikasi status kemiskinan rumah tangga dengan support vector machines (SVM) dan classification and regression trees (CART) menggunakan GUI R (studi kasus di Kabupaten Wonosobo tahun 2018). Jurnal Gaussian, 9(4), 525–534.

Oktavia, A. D. (2018). Faktor-faktor yang berpengaruh dalam mendapatkan pekerjaan bagi lulusan statistika IPB dengan menggunakan metode CHAID (studi kasus: alumni departemen statistika IPB angkatan 48-50). Skripsi, Departemen Statistika, Institut Pertanian Bogor, Bogor.

Sartono, B., & Syafitri, U. D. (2010). Metode pohon gabungan : solusi pilihan untuk mengatasi kelemahan pohon regresi dan klasifikasi tunggal. Forum Statistika Dan Komputasi, 15(1), 1–7.

Sulviana, V. (2018). Penggunaan metode CHAID (chi-squared automatic interaction detection) pada segmentasi tren penjualan berbagai jenis minuman ringan di Indonesia. Skripsi, Departemen Statistika, Institut Pertanian Bogor, Bogor.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. Handbook of Statistics, 24(04), 303–329. https://doi.org/10.1016/S0169-7161(04)24011-1

Utami, T. P. (2019). Penerapan metode random forest dalam menentukan status istitaah kesehatan jemaah haji (studi kasus: jemaah haji di Kecamatan Plered, Kabupaten Purwakarta). Skripsi, Departemen Statistika, Institut Pertanian Bogor, Bogor.

Yanthy, M. (2013). Penentuan karakteristik kelancaran pembayaran kartu kredit menggunakan metode CHAID. Skripsi, Departemen Statistika, Institut Pertanian Bogor, Bogor.