JSA
Jurnal Statistika dan Aplikasinya

Volume 6 Issue 1,
June 2022

e-ISSN: 2620-8369

Received: 20 May 2022
Revised: 13 June 2022
Accepted: 26 June 2022
Published: 30 June 2022

Penerapan Metode SMOTE CHAID dalam Klasifikasi Tuberkulosis *Relapse*

Vera Maya Santi[1, a)], Lina Nafisah[1, b)], Qorry Meidianingsih[2, c)]

[1]*Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta, Jl. Rawamangun muka, Kota Jakarta Timur, DKI Jakarta, 13220.*
[2]*Program Studi Pendidikan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta, Jl. Rawamangun muka, Kota Jakarta Timur, DKI Jakarta, 13220.*

Email: [a)] vmsanti@unj.ac.id, [b)] linanafisah24@gmail.com, [c)]qorrymeidianingsih@unj.ac.id

Abstract

DKI Jakarta Province is one of the provinces with the highest tuberculosis cases, and a person's chance of contracting tuberculosis is the greatest among other provinces. Tuberculosis can be cured with regular treatment within a certain period of time, but after recovering, some tuberculosis sufferers may *relapse* so that it can cause new problems. This study aims to build a classification model and determine which factors influence tuberculosis *relapse* using the CHAID method. SMOTE with majority under sampling is applied as a solution to deal with the problem of patient categories (*relapse* and non-*relapse*) who have an unbalanced number of observations. Based on the CHAID classification tree, the results show that the factors that influence *relapse* in tuberculosis patients include the type of diagnosis, age, gender, and place of residence. In addition, the application of SMOTE can improve the performance of the CHAID classification tree in classifying patients based on their categories. These results were indicated by an increase in the value of sensitivity to 26,667 compared to the performance of CHAID without SMOTE. Based on these results, the SMOTE CHAID classification model has better performance than CHAID.

Keywords: CHAID, tuberculosis, *relapse*, imbalanced class, SMOTE.

Abstrak

Provinsi DKI Jakarta merupakan salah satu provinsi dengan kasus tuberkulosis yang tinggi dan peluang seseorang untuk terkena tuberkulosis merupakan peluang paling besar diantara provinsi lainnya. Tuberkulosis dapat disembuhkan dengan pengobatan rutin dalam jangka waktu tertentu, tetapi setelah sembuh sebagian penderita tuberkulosis dapat mengalami *relaps* (kambuh) sehingga dapat menimbulkan masalah baru. Penelitian ini bertujuan untuk membangun model klasifikasi tuberkulosis *relapse* menggunakan metode klasifikasi SMOTE CHAID dan mengetahui faktor - faktor apa saja yang mempengaruhi kejadian *relapse* pada pasien tuberkulosis. SMOTE *with majority undersampling* diterapkan sebagai solusi dalam menangani

permasalahan kategori pasien (*relapse* dan tidak *relapse*) yang memiliki jumlah amatan tidak seimbang. Berdasarkan pohon klasifikasi CHAID, diperoleh hasil bahwa faktor – faktor yang mempengaruhi *relapse* pada pasien tuberkulosis di antaranya tipe diagnosis, umur, jenis kelamin, dan asal tempat tinggal. Selain itu, penerapan SMOTE mampu meningkatkan performa pohon klasifikasi CHAID dalam mengklasifikasikan pasien berdasarkan kategorinya. Hasil tersebut ditunjukkan dengan peningkatan sensitivitas menjadi 26,667 dibandingkan dengan performa CHAID tanpa adanya SMOTE. Berdasarkan hasil tersebut, model klasifikasi SMOTE CHAID memiliki performa lebih baik dibandingkan dengan CHAID tanpa SMOTE.

Kata-kata kunci: CHAID, tuberkulosis, *relapse*, *imbalanced class,* SMOTE.

PENDAHULUAN

Indonesia masih merupakan daerah endemik penyakit tuberkulosis. Hal ini terbukti pada tahun 2018 Indonesia berada di urutan ketiga dari 30 *high burden countries* terhadap tuberkulosis paru dengan jumlah kasus sebanyak 570.289 (*World Health Organization*, 2019). Upaya penanggulangan maupun pencegahan telah diupayakan oleh pemerintah Indonesia untuk menurunkan angka kesakitan dan kematian penderita tuberkulosis, tetapi masih belum berhasil sepenuhnya dalam menyelesaikan masalah. Penyakit yang mudah menular ini sebenarnya dapat disembuhkan dengan pengobatan rutin dalam jangka waktu tertentu, tetapi setelah sembuh sebagian penderita tuberkulosis dapat mengalami *relapse* (kambuh). Adanya kejadian kasus *relapse* ini dapat menimbulkan masalah baru karena meningkatkan kemungkinan resistensi obat anti tuberkulosis (Naomi et. al, 2016). Salah satu upaya untuk menanggulangi tuberkulosis paru *relapse* di Indonesia adalah dengan melakukan penelitian mengenai karakteristik penderita tuberkulosis *relapse*.

Beberapa metode analisis statistika dapat diterapkan dalam menentukan karakteristik atau klasifikasi dari suatu peubah. Salah satunya menggunakan metode *Chi Square Automatic Interaction Detection* (CHAID) yang umumnya dikenal sebagai metode pohon klasifikasi (*Classification tree methods*). Tetapi sekumpulan data yang akan diklasifikasikan sering mengalami ketidakseimbangan kelas. Ketidakseimbangan kelas terjadi ketika data yang dimiliki pada masing-masing kelas peubah respon pada suatu kelas lebih banyak (mayoritas) dibandingkan kelas lainnya (minoritas). Apabila klasifikasi ini tetap dilakukan pada data tidak seimbang maka akan menyebabkan terjadinya *overfitting* atau cenderung mengabaikan kelas minoritas sehingga dapat berpengaruh buruk terhadap performa algoritma klasifikasi (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

Metode *oversampling* yang berlebihan dapat menyebabkan *overfitting* sedangkan *undersampling* yang berlebihan dapat berpengaruh pada hilangnya beberapa informasi penting yang terdapat pada data set (Napierała, 2012). Untuk menghindari kemungkinan *overfitting* dalam *oversampling*, salah satu metode *oversampling* yang paling dikenal yaitu *Synthetic Minority Oversampling Technique* (SMOTE) dimana metode ini dapat menambah jumlah data kelas minoritas agar setara dengan kelas mayor dengan cara membangkitkan data buatan berdasarkan k-tetangga terdekat (*k-nearest neighbor*). Meskipun begitu, terdapat cara untuk meningkatkan performa metode *oversampling* dimana salah satunya adalah dengan menggabungkan metode *undersampling* sebagai metode pembersih (Choirunnisa, 2019). Penggabungan dengan metode *undersampling* ini diharapkan agar data yang diproses menjadi lebih bersih dan terhindar dari *noise* sehingga mampu meningkatkan kemampuan metode *oversampling* dalam membuat data sintetik. Dalam penelitian ini, sebelum dilakukan penerapan metode CHAID akan dilakukan penerapan metode SMOTE *with majority undersampling* yang diharapkan dapat menghasilkan performa klasifikasi yang lebih baik.

METODOLOGI

Bahan dan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yaitu rekam medis pasien tuberkulosis yang berobat di Puskesmas Kecamatan Setiabudi Kota madya Jakarta Selatan, Provinsi DKI Jakarta selama bulan Januari 2017 – Maret 2021. Berdasarkan data tersebut, terdapat 650 responden yang terbagi ke dalam kategori *relapse* (77 penderita atau sekitar 11,8 %) dan tidak *relapse* (573 penderita atau sekitar 88,2 %). Pemilihan variabel yang digunakan dalam penelitian ini mengacu pada penelitian Cruz, A. P. Dela. (2018) dan W. S. Fitri & Munir (2014). Variabel penelitian yang digunakan terdiri dari satu peubah respon (Y) dan lima peubah penjelas (X) seperti yang ditampilkan pada TABEL 1.

TABEL 1. Variabel Penelitian

Notasi	Nama Variabel	Tipe	Kategori
Y	Kategori Pasien	Nominal	1: Pasien *relapse* 2: Pasien Tidak *relapse*
X_1	Umur	Nominal	1: < 50 2: > 50
X_2	Jenis Kelamin	Nominal	1: Laki-laki 2: Perempuan
X_3	Asal kelurahan Pasien	Nominal	1: Kel. Setiabudi 2: Kel. Guntur 3: Kel. Karet Semanggi 4: Kel. Karet Kuningan 5: Kel. Kuningan Timur 6: Kel. Menteng Atas 7: Kel. Pasar Manggis 8: Di luar Kec.Setiabudi
X_4	Tipe diagnosis tuberkulosis	Nominal	1: Terkonfirmasi bakterologis 2: Terdiagnosis klinis
X_5	Klasifikasi Pasien tuberkulosis	Nominal	1: Paru 2: Extra paru

Metode Penelitian

Tahapan analisis data yang dilakukan dalam penelitian ini sebagai berikut:

1. Melakukan pra-pemrosesan data dengan memberi label pada pasien, melakukan seleksi peubah, *cleaning* data dan pengkategorian data. Seleksi peubah dan *cleaning* data dilakukan dengan tujuan untuk mendapatkan data yang siap digunakan dalam penelitian.
2. Membuat diagram frekuensi antara pasien *relapse* dan tidak *relapse* dan melakukan eksplorasi data dengan menggunakan diagram pie dalam membandingkan data untuk melihat karakteristik responden.
3. Pembagian data *training* dan data *testing* dengan teknik pengambilan sampel acak sederhana. Perbandingan data *training* dan data *testing* sebesar 80% : 20% (Suantari, 2020).
4. Membentuk model klasifikasi
   Terdapat dua model yang akan dibentuk dalam proses ini dengan menggunakan *R Studio*, yaitu:
   a. Model CHAID tanpa penerapan SMOTE
      - Melakukan proses matematis analisis CHAID pada data *training* untuk melihat struktur data antara peubah penjelas dengan peubah respon serta evaluasi model. Proses ini akan

menerapkan 3 langkah analisis CHAID, yaitu langkah Penggabungan, Pemisahan, dan Pemberhentian.

- Dalam langkah penggabungan, akan mulai diterapkan uji *chi-square* dan pengali *Bonferroni* sebagai pengoreksinya. Pada langkah penggabungan sebagian besar proses akan menggunakan uji *chi-square*.
- Kemudian pada langkah pemisahan akan dilakukan pemilihan peubah penjelas yang akan digunakan sebagai *split node* (pemisah simpul) yang terbaik.
- Kemudian dilakukan iterasi pada kedua langkah tersebut, dan proses iterasi akan berhenti apabila sudah tidak ada lagi peubah penjelas yang tersisa untuk diuji hubungannya dengan peubah respons, atau juga apabila terbentuknya simpul pada diagram pohon telah memenuhi batasan yang ditentukan oleh peneliti. Proses ini disebut dengan proses *stopping time*.
- Kemudian model yang dihasilkan disebut model 1.

b. Model CHAID dengan penerapan SMOTE *with majority undersampling*

- Menerapkan metode SMOTE untuk membangkitkan data sintesis/buatan lalu dilanjutkan ke tahap CHAID. Tahapan di dalam proses SMOTE adalah sebagai berikut :

i. Menentukan k-tetangga terdekat (pada penelitian ini digunakan k=5) dan menghitung jarak antara contoh dan tetangga terdekatnya menggunakan modus yang perhitungannya menggunakan *Value Distance Metric (VDM)* sebagai berikut:

$$\Delta(X,Y) = w_A w_B \sum_{i=1}^{p} \delta(x_1, y_2)^r \qquad (1)$$

Keterangan:

$\Delta(X,Y)$: jarak antara amatan X dan Y
$w_A w_B$: bobot amatan (dapat diabaikan)
p : banyaknya peubah penjelas
r : bernilai 1 (jarak Manhattan) atau 2 (jarak Euclidean)
$\delta(x_1, y_2)$: jarak antar amatan X dan Y untuk setiap peubah penjelas.

ii. Membangkitkan data buatan dengan tahapan sebagai berikut:

Data numerik

a) Berdasarkan 5 tetangga terdekat yang dihasilkan dari langkah sebelumnya, pilih satu secara acak.
b) Hitung selisih antara data amatan dengan tetangga terdekat yang terpilih.
c) Kalikan selisih yang diperoleh dari langkah sebelumnya dengan angka acak antara 0 dan 1.
d) Nilai yang diperoleh di tambahkan dengan nilai data amatan asli. Hasil tersebut merupakan data buatan yang dibangkitkan.
e) Lakukan sebanyak *n* kali sesuai dengan jumlah persentase *oversampling* yang digunakan. Pada penelitian ini digunakan persentase *oversampling* sebesar 400, 500, 600.
f) Masukkan persentase *undersampling* hingga kelas mayoritas dan minor seimbang.

Data kategorik

a) Berdasarkan 5 tetangga terdekat yang dihasilkan dari langkah sebelumnya, data buatan yang dihasilkan merupakan nilai modus dari kelima tetangga terdekatnya. Jika terdapat nilai modus yang sama maka pilih secara acak.
b) Lakukan sebanyak *n* kali sesuai dengan jumlah persentase *oversampling* yang digunakan. Pada penelitian ini digunakan persentase *oversampling* sebanyak 400, 500, 600.
c) Masukkan persen *undersampling* hingga kelas mayoritas dan minor seimbang.

- Kemudian model yang dihasilkan disebut model 2.

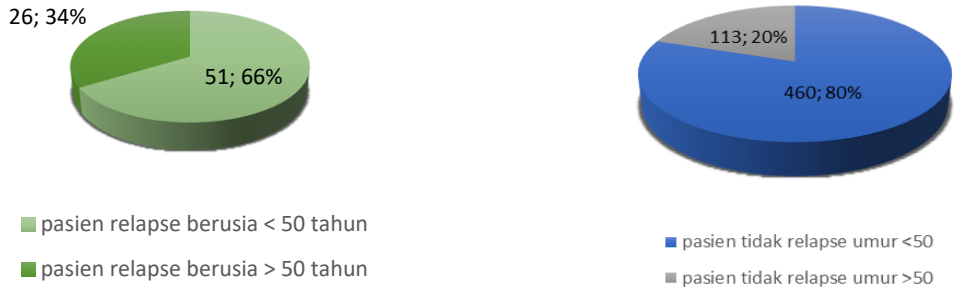
5. Melakukan prediksi dengan data testing berdasarkan metode klasifikasi CHAID dan SMOTE CHAID.

6. Mengevaluasi model CHAID dan SMOTE dengan menghitung nilai akurasi, sensitivas, dan spesifisitas.
7. Membandingkan performa CHAID dan SMOTE CHAID.
8. Interpretasi model terbaik.

HASIL DAN PEMBAHASAN

Eksplorasi Data

Eksplorasi data pada penelitian ini menggunakan diagram pie yang bertujuan membandingkan data untuk melihat karakteristik responden. Hasil diagram pie dari masing masing hubungan peubah respon dengan peubah penjelas disajikan pada gambar dibawah ini.



(a) **(b)**

GAMBAR 1. Persentase hubungan umur berdasarkan kategori pasien tuberkulosis: (a) *relapse*, (b) tidak *relapse*

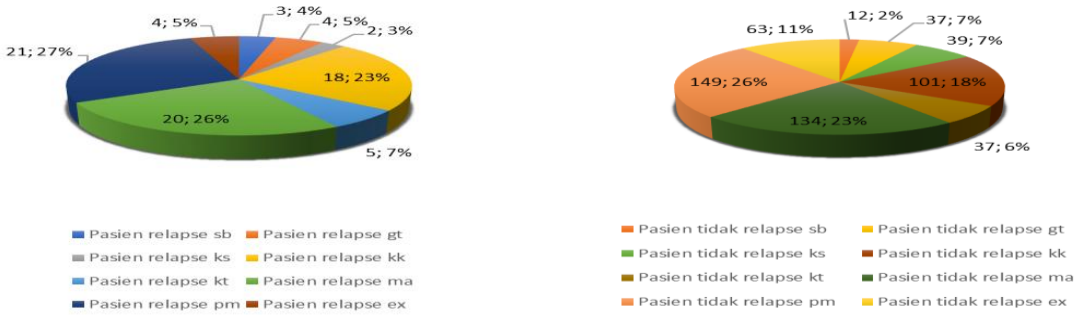
Berdasarkan Gambar 1 (a) persentase pasien tuberkulosis *relapse* berusia <50 tahun memiliki persentase yang lebih besar dibandingkan dengan pasien tuberkulosis *relapse* berusia >50 tahun. Menurut Gambar 1 (b) pasien tuberkulosis tidak *relapse* berusia <50 tahun memiliki persentase lebih besar juga daripada pasien tuberkulosis tidak *relapse* berusia >50 tahun. Di antara semua kategori, persentase pasien berstatus tidak *relapse* berusia <50 tahun memiliki jumlah pasien tuberkulosis terbanyak dengan jumlah 460 pasien. Dengan melihat keunggulan pasien berusia <50 tahun di setiap kategori, artinya pasien tuberkulosis di Puskesmas Kecamatan Setiabudi paling banyak yaitu berusia <50 tahun.

(a) **(b)**

GAMBAR 2. Persentase hubungan jenis kelamin berdasarkan kategori pasien tuberkulosis: (a) *relapse*, (b) tidak *relapse*

Berdasarkan Gambar 2 (a) persentase pasien tuberkulosis *relapse* yang berjenis kelamin laki - laki memiliki persentase yang lebih besar dibandingkan dengan pasien tuberkulosis *relapse* yang

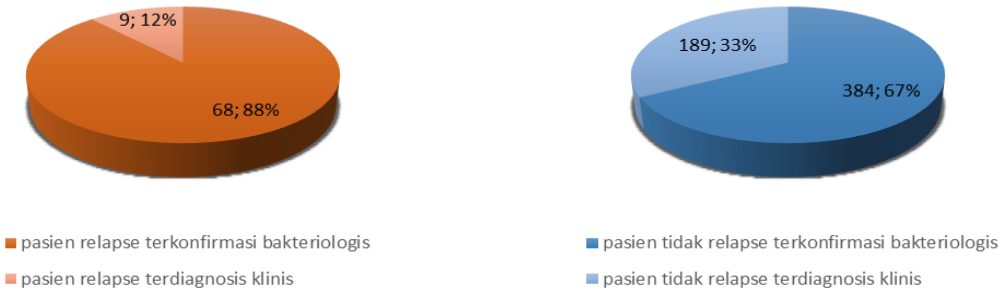
berjenis kelamin perempuan. Menurut Gambar 2 (b) pasien tuberkulosis tidak *relapse* yang berjenis kelamin laki - laki memiliki persentase lebih besar juga daripada pasien tuberkulosis tidak *relapse* yang berjenis kelamin perempuan. Di antara semua kategori, persentase pasien berstatus tidak *relapse* yang berjenis kelamin laki - laki memiliki jumlah pasien tuberkulosis terbanyak dengan 353 pasien. Dengan melihat keunggulan pasien berjenis kelamin laki - laki di setiap kategori, artinya pasien tuberkulosis di Puskesmas Kecamatan Setiabudi paling banyak yaitu berjenis kelamin laki - laki.



(a) **(b)**

GAMBAR 3. Persentase hubungan asal tempat tinggal berdasarkan kategori pasien tuberkulosis: (a) *relapse*, (b) tidak *relapse*

Berdasarkan Gambar 3 (a) persentase pasien tuberkulosis *relapse* yang bertempat tinggal di kelurahan Pasar Manggis memiliki persentase paling besar dan persentase pasien tuberkulosis *relapse* yang bertempat tinggal di kelurahan Karet Semanggi memiliki persentase paling kecil. Menurut Gambar 3 (b) pasien tuberkulosis tidak *relapse* yang bertempat tinggal di kelurahan Pasar Manggis memiliki persentase paling besar dan persentase pasien tuberkulosis tidak *relapse* yang bertempat tinggal di kelurahan Setiabudi memiliki persentase paling kecil juga. Diantara semua kategori, persentase pasien berstatus tidak *relapse* yang bertempat tinggal di Kelurahan Pasar Manggis memiliki jumlah pasien tuberkulosis terbanyak dengan 149 pasien. Dengan melihat keunggulan pasien dari kelurahan Pasar Manggis, artinya pasien tuberkulosis di Puskesmas Kecamatan Setiabudi paling banyak yaitu bertempat tinggal di Kelurahan Pasar Manggis.

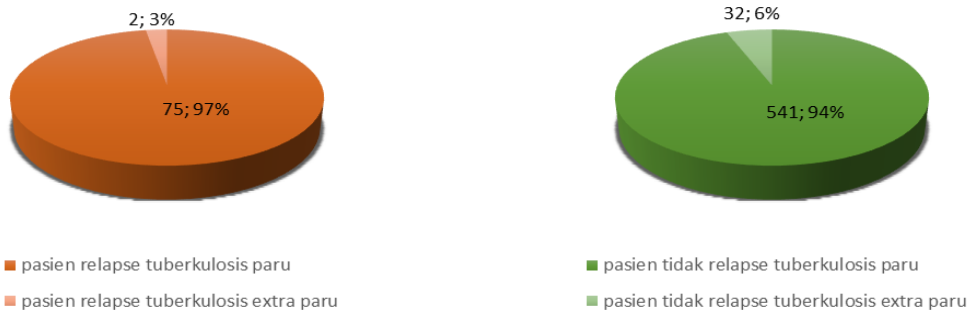


(a) **(b)**

GAMBAR 4. Persentase hubungan tipe diagnosis berdasarkan kategori pasien tuberkulosis: (a) *relapse*, (b) tidak *relapse*

Berdasarkan Gambar 4 (a) persentase pasien tuberkulosis *relapse* yang terkonfirmasi bakteriologis memiliki persentase yang lebih besar dibandingkan dengan pasien tuberkulosis *relapse* yang terdiagnosis klinis. Menurut Gambar 4 (b) pasien tuberkulosis tidak *relapse* yang terkonfirmasi bakteriologis memiliki persentase lebih besar juga daripada pasien tuberkulosis tidak *relapse* yang terdisgnosis klinis. Di antara semua kategori, persentase pasien berstatus tidak *relapse* yang terkonfirmasi bakteriologis memiliki jumlah pasien tuberkulosis terbanyak dengan jumlah 384 pasien.

Dengan melihat keunggulan pasien terkonfirmasi bakteriologis disetiap kategori, artinya pasien tuberkulosis di Puskesmas Kecamatan Setiabudi paling banyak yaitu yang terkonfirmasi bakteriologis.



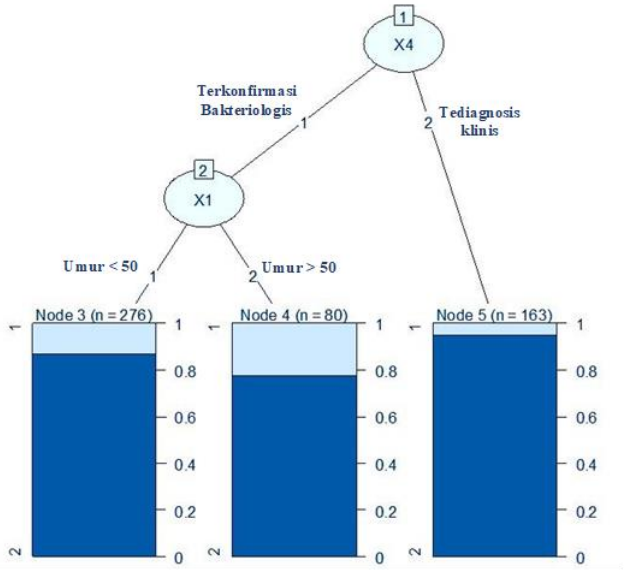
(a) **(b)**

GAMBAR 5. Persentase hubungan klasifikasi pasien berdasarkan kategori pasien tuberkulosis: (a) *relapse*, (b) tidak *relapse*

Berdasarkan Gambar 5 (a) persentase pasien *relapse* tuberkulosis paru memiliki persentase yang lebih besar dibandingkan dengan pasien *relapse* tuberkulosis extra paru. Menurut Gambar 5 (b) pasien tidak *relapse* tuberkulosis paru memiliki persentase lebih besar juga daripada pasien tidak *relapse* tuberkulosis extra paru. Di antara semua kategori, persentase pasien tidak *relapse* yang tuberkulosis extra paru memiliki jumlah pasien terbanyak dengan 541 penderita. Dengan melihat keunggulan pasien tuberkulosis paru disetiap kategori, artinya pasien tuberkulosis di Puskesmas Kecamatan Setiabudi paling banyak yaitu tuberkulosis paru.

Klasifikasi CHAID Tanpa SMOTE

Segmentasi yang dihasilkan oleh analisis CHAID pada keseluruhan data training dapat dilihat dari diagram pohon klasifikasi pada GAMBAR 6 sebagai berikut:



GAMBAR 6. Diagram pohon hasil analisis CHAID pada keseluruhan data

Gambar 6 menunjukkan bahwa analisis CHAID menghasilkan tiga node akhir yaitu node 3, node 4 dan node 5. Hasil klasifikasi pada node 3 adalah pasien yang terkonfirmasi bakteriologis dan

berusia <50 tahun berjumlah 276 penderita dengan kemungkinan akan terkena *relapse* sekitar 13%. Hasil klasifikasi pada node 4 adalah pasien yang terkonfirmasi bakteriologis dan berusia >50 tahun berjumlah 80 penderita dengan kemungkinan akan terkena *relapse* sekitar 22,5%. Untuk hasil klasifikasi node 5 adalah pasien yang terdiagnosis klinis berjumlah 163 dengan kemungkinan akan terkena *relapse* sekitar 4,9%. Hal ini menunjukkan bahwa pasien yang terkonfirmasi secara bakteriologis berusia <50 tahun akan lebih banyak yang mengalami *relapse* dibandingkan jalur masuk lainnya.

Bangkitan Data Sintesis Dengan SMOTE

Tahap pertama dalam penanganan ketidakseimbangan kelas menggunakan metode SMOTE *with majority undersampling* yaitu menentukan persentase *oversampling* dan *undersampling* yang akan diterapkan pada data training sebagai berikut:

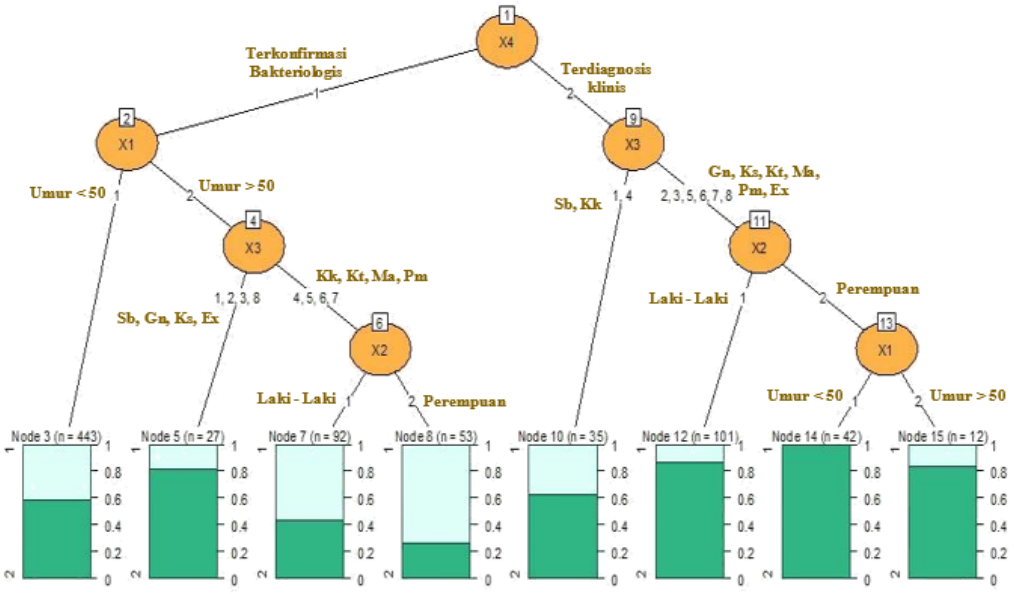
TABEL 2. Rincian nilai akurasi, sensitivitas, dan spesifisitas dalam pembangkitan data

No.	Perc. Over	Perc. Under	∑Data Minor	∑Data Mayor	Akurasi	Spesifisitas	Sensitivitas
1	400%	125%	310	310	36,923	29,565	93,333
2	400%	200%	310	496	76,153	82,608	26,667
3	500%	120%	372	372	43,846	37,391	93,333
4	500%	200%	372	620	70	74,782	33,3
5	600%	116%	434	431	37,692	32,173	80
6	600%	200%	434	744	74,615	81,739	20

Kombinasi pertama dan ketiga memiliki nilai sensitivitas rataan yang lebih baik tetapi mempunyai nilai akurasi yang paling kecil daripada kombinasi lainnya. Berdasarkan pertimbangan nilai akurasi dan spesifisitas rataan, kombinasi persentase *oversampling* dan *undersampling* yang digunakan yaitu kombinasi ke-2 atau dengan persentase *oversampling* sebesar 400% dan persentase *undersampling* sebesar 200%. Gugus data baru yang dihasilkan berdasarkan kombinasi tersebut selanjutnya dianalisis untuk menentukan faktor-faktor yang memengaruhi tuberkulosis *relapse*.

Klasifikasi CHAID Menggunakan SMOTE

Hasil analisis CHAID menunjukkan bahwa empat peubah penjelas berasosiasi signifikan terhadap peubah kategori pasien seperti di Gambar 7. Pada Gambar 7 menunjukkan bahwa peubah penjelas yang dimaksud adalah tipe diagnosis, asal kelurahan pasien, jenis kelamin, dan umur. Peubah penjelas yang tidak berasosiasi signifikan terhadap kategori pasien tuberkulosis adalah peubah klasifikasi pasien. Analisis CHAID menghasilkan delapan node akhir. Di node pertama terbagi menjadi 2 bagian yaitu pasien yang terkonfirmasi bakteriologis dan terdiagnosis klinis.



GAMBAR 7. Diagram pohon hasil analisis CHAID SMOTE pada keseluruhan data

TABEL 3 berikut mendeskripsikan hasil diagram pohon di atas.

TABEL 3. Nilai *gains* untuk node akhir pada data bangkitan SMOTE

Node	Karakteristik Pasien	Node		Gain	
		N	Persentase	N	Persentase
3	Pasien yang terkonfirmasi bakteriologis dan berusia <50 tahun	443	54,96%	185	59,68%
5	Pasien terkonfirmasi bakteriologis dan berusia >50 tahun yang bertempat tinggal di Kel. Setiabudi, Kel. Guntur, Kel. Karet Semanggi dan di luar Kecamatan Setiabudi	27	3,35%	5	1,61%
7	Pasien terkonfirmasi bakteriologis dan berusia >50 tahun yang bertempat tinggal di Kel. Karet Kuningan, Kel. Kuningan Timur, Kel. Menteng Atas, Kel. Pasar Manggis dan berjenis kelamin laki-laki	92	11,41%	52	16,77%
8	Pasien yang terkonfirmasi bakteriologis dan berusia >50 tahun yang bertempat tinggal di Kel. Karet Kuningan, Kel. Kuningan Timur, Kel. Menteng Atas, Kel. Pasar Manggis dan berjenis kelamin perempuan	53	6,58%	39	12.58%
10	Pasien yang terdiagnosis klinis dan bertempat tinggal di Kel. Setiabudi, Kel. Karet Kuningan	35	4,34%	13	4,19%
12	Pasien yang terdiagnosis klinis dan bertempat tinggal di Kel. Guntur, Kel. Karet Semanggi, Kel. Kuningan Timur, Kel. Menteng Atas, Kel. Pasar Manggis atau di luar kecamatan Setiabudi dan berjenis kelamin laki-laki	101	12,53%	14	4,51%

14	Pasien yang terdiagnosis klinis dan bertempat tinggal di Kel. Guntur, Kel. Karet Semanggi, Kel. Kuningan Timur, Kel. Menteng Atas, Kel. Pasar Manggis atau di luar kecamatan Setiabudi dan berjenis kelamin perempuan dengan umur <50 tahun	42	5,21%	0	0%
15	Pasien yang terdiagnosis klinis dan bertempat tinggal di Kel. Guntur, Kel. Karet Semanggi, Kel. Kuningan Timur, Kel. Menteng Atas, Kel. Pasar Manggis atau di luar kecamatan Setiabudi dan berjenis kelamin perempuan	12	1,49	2	0,65%

Perbandingan Performa Hasil klasifikasi

TABEL 4 menampilkan hasil performa klasifikasi rataan analisis CHAID pada data testing sebelum dan sesudah diterapkannya metode SMOTE sebagai berikut:

TABEL 4. Perbandingan Hasil Performa CHAID dengan CHAID SMOTE

Nilai	Hasil Performa CHAID	
	Tanpa SMOTE	Sesudah SMOTE
Akurasi	88,461	76,153
Sensitivitas	0	26,667
Spesifisitas	100	82,608

Performa klasifikasi pada data sebelum diterapkannya metode SMOTE menghasilkan nilai akurasi dan spesifisitas rataan yang tinggi berturut turut sebesar 88,461 dan 100. Nilai spesifisitas rataan sebesar 100 menunjukkan bahwa pasien tuberkulosis yang berstatus tidak *relapse* sudah dapat diklasifikasikan dengan benar. Hal ini berbanding terbalik dengan nilai sensitivitas rataan yang sangat rendah sebesar 0. Kondisi ini menunjukkan bahwa seluruh pasien tuberkulosis yang berstatus *relapse* salah diklasifikasikan sebagai pasien tuberkulosis tidak *relapse*. Performa klasifikasi pada data sesudah diterapkannya metode SMOTE menghasilkan nilai akurasi, sensitivitas, dan spesifisitas rataan berturut-turut sebesar 76,153; 26,667; dan 82,608. Tabel 3 menunjukkan bahwa nilai sensitivitas rataan mengalami kenaikan dari 0 menjadi 26,667 pada data yang telah diterapkannya SMOTE. Dalam kasus tidak seimbang, akurasi tidak tepat digunakan sebagai ukuran performa klasifikasi. Maka tidak masalah sekalipun nilai akurasi menurun pada SMOTE CHAID (Fitriani, Yasin, & Tarno, 2021). Jika sensitivitas meningkat maka spesifisitas menurun (Rianto & Wahono, 2015). Hal ini berarti analisis CHAID pada data bangkitan SMOTE lebih baik dalam mengklasifikasi pasien yang berstatus *relapse* dibandingkan analisis CHAID pada data sebelum diterapkan SMOTE.

KESIMPULAN DAN SARAN

Berdasarkan analisis dan hasil yang sudah dibahas sebelumnya, penerapan SMOTE pada data pasien tuberkulosis *relapse* dan tidak *relapse* di Puskesmas Kecamatan Setiabudi pada bulan Januari 2017 – Maret 2021 ini mampu meningkatkan kemampuan metode klasifikasi CHAID. Hal ini dapat dilihat dari nilai akurasi, sensitivitas, dan spesifisitas rataan menjadi 76,153; 26,667; dan 82,608. Peubah penjelas yang berasosiasi signifikan terhadap kategori pasien tuberkulosis dengan

menggunakan analisis CHAID pada data bangkitan SMOTE adalah tipe diagnosis, umur, tempat tinggal dan jenis kelamin.

Saran yang dapat diberikan yaitu pada penerapan metode penanganan kelas tidak seimbang *bagging* atau *bootstrap aggregating* dapat dijadikan sebagai alternatif bersamaan dengan metode klasifikasi lainnya, seperti CART, regresi logistik, *random forest*, *support vector machine* dan lainnya.

REFERENSI

World Health Organization. (2019). *Global Tuberculosis Report 2018.* Retrieved from https://www.who.int/Tuberkulosis/publications/global_report/en/

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *Journal of Artificial Intelligence Research*, *16*(Sept. 28), 321–357. snopes.com: Two-Striped Telamonia Spider.

Napierała, K. (2012). *Improving Rule Classifiers For Imbalanced Data*. Retrieved from http://www.cs.put.poznan.pl/knapierala/misc/pssi2013konkursNapieralarozprawa.pdf

Cruz, A. P. Dela. (2018). *Predicting the Relapse Category in Patients with Tuberculosis: A Chi-Square Automatic Interaction Detector (CHAID) Decision Tree Analysis.* Open Journal of Social Sciences, 06(12), 29–36. https://doi.org/10.4236/jss.2018.612003

Fitri, W. S., & Munir, S. M. (2014). *Karakteristik Penderita Tuberkulosis Paru Relapse Yang berobat di Poliklinik Paru Rumah Sakit Umum Daerah Arifin Achmad Provinsi Riau Tahun 2012-2013*. 1(c), 1–43.

Fitriani, R. D., Yasin, H., & Tarno, T. (2021). *Penanganan Klasifikasi Kelas Data Tidak Seimbang Dengan Random Oversampling Pada Naive Bayes* (Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal). *Jurnal Gaussian*, *10*(1), 11–20. https://doi.org/10.14710/j.gauss.v10i1.30243

Rianto, H., & Wahono, R. S. (2015). Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software. IlmuKomputer.Com *Journal of Software Engineering*, 1(1), 46–53.

Suantari, ni gusti ayu putu puter. (2020). *Penerapan smote pada metode chaid untuk mengklasifikasi tingkat loyalitas pelanggan*. Institut Pertanian Bogor.

Naomi, D. A., Dilangga, P., Ramadhian, M. R., & Marlina, N. (2016). Penatalaksanaan Tuberkulosis Paru Kasus Kambuh pada Wanita Usia 32 Tahun di Wilayah Rajabasa. *J Medula Unila*, 6, 20–27.

Choirunnisa, S. (2019). *Metode Hibrida Oversampling Dan Ketidakseimbangan Data Kegagalan*.