

Komparasi Performa Algoritma ID3, C4.5, CHAID Dalam Profiling Tersangka Kasus Narkoba Di Jawa Barat

Asep Permana
Universitas Budi Luhur
Jakarta Selatan, Indonesia
asperscript@gmail.com

Agustia Hananto
Universitas Budi Luhur
Jakarta Selatan, Indonesia
agustia.hananto@gmail.com

Sigit Budi Nugroho
Universitas Budi Luhur
Jakarta Selatan, Indonesia
sigitsoeharboen@gmail.com

Arief Wibowo
Universitas Budi Luhur
Jakarta Selatan, Indonesia
arief.wibowo@budiluhur.ac.id

Abstract— Strategi dalam pengendalian kejahatan narkoba tidak bisa didasari melalui intervensi berupa larangan saja, akan tetapi memahami pola kejahatan narkoba akan menjadi lebih baik dalam pengendaliannya, terutama untuk pencegahan dan kewaspadaan. Dengan menggunakan teknik *data mining*, *insigth* terhadap kejahatan narkoba dapat didefinisikan dengan baik, salah satunya dengan algoritma klasifikasi *decision tree*. Ada beberapa algoritma dengan basis *decision tree*, seperti ID3, C4.5, dan CHAID. Dengan banyaknya pilihan algoritma, maka metode perbandingan algoritma dipilih oleh penulis dalam penelitian ini, tujuannya untuk menentukan algoritma yang paling tepat dalam menganalisis data. Performa akurasi menjadi tolok ukur dalam perbandingannya, dan hasil dari perbandingan yang dilakukan, penulis memilih algoritma CHAID sebagai pemodelannya dengan nilai akurasi sebesar 73.89%, sedangkan yang lainnya masing-masing memiliki nilai akurasi sebesar 70.14% untuk ID3 dan 72.44% untuk C4.5.

Kata kunci — C4.5, CHAID, data mining, ID3

I. PENDAHULUAN

Segala jenis kejahatan akan memberikan dampak yang buruk, tidak terkecuali kejahatan terkait penyalahgunaan narkoba, selain membahayakan kesehatan, kejahatan terkait narkoba beberapa diantaranya dapat memicu kejahatan lain seperti penipuan, perampokan, pencurian, hingga kekerasan secara fisik untuk mendapatkan uang agar dapat membeli narkoba. Konsekuensi negatif yang timbul akibat penggunaan narkoba, paling jelas terlihat pada individu pelakunya, kesehatan semakin memburuk dan rentan tertular oleh penyakit berbahaya seperti HIV/AIDS, sampai akhirnya berdampak pada kematian. Akibatnya pengaruh dari kejahatan narkoba memberikan dampak buruk yang luas bagi kehidupan sosial [1]. Pada tahun 2018 Badan Narkotika Nasional (BNN) mencatat penyalahgunaan narkoba mencapai 2.29 juta orang dari kalangan pelajar di seluruh Indonesia dan terjadi peningkatan sebesar 24 hingga 28% di tahun 2019 [2]. Melihat kondisi yang seperti ini, penelitian terhadap kejahatan penyalahgunaan narkoba menjadi sangat perlu untuk dilakukan, agar dapat meningkatkan kewaspadaan kita di masa yang akan datang. *Data mining* dengan metode klasifikasi dapat memprediksi pola dari pelaku, info demografis dapat berperan dalam memprediksi pelaku penyalahgunaan narkoba [3] atau lebih dikenal dengan istilah *profiling*.

Profiling tersangka pelaku kejahatan merupakan pendekatan yang umum dilakukan dalam proses investigasi dengan tujuan agar mendapatkan ciri-ciri dari pelaku yang bersangkutan. Metode klasifikasi pada *data mining* telah banyak digunakan oleh para peneliti sebagai teknik untuk melakukan *profiling*. Pemilihan metode dan algoritma yang tepat, menjadi kunci dalam melakukan *profiling* [4], sehingga algoritma yang digunakan memiliki performa yang baik dari tingkat akurasi. Ada banyak algoritma yang dapat digunakan pada metode klasifikasi, namun permasalahannya adalah bagaimana kita dapat menentukan algoritma terbaik untuk *profiling* tersangka penyalahgunaan narkoba? Karena memang tidak ada standar algoritma yang baku [5] untuk penerapannya, semuanya tergantung data yang akan dijadikan penelitian. Maka dari itu, tujuan dari penelitian ini adalah penerapan metode komparasi terhadap performa algoritma dari ID3, C4.5, dan CHAID untuk menentukan algoritma yang tepat dalam menganalisis *profiling* pelaku tindak kejahatan penyalahgunaan narkoba.

II. TINJAUAN PUSTAKA

Analisis mengenai tindak kejahatan menggunakan pendekatan *machine learning* sudah banyak dilakukan sebelumnya, seperti menggunakan algoritma Naïve Bayes [6] untuk analisis pola kejahatan pada media sosial facebook. Pendekatan dengan algoritma K-Means [7] untuk klusterisasi tingkat kejahatan di daerah Pematangsiantar. Sedangkan penelitian sebelumnya tentang pelaku tindak kejahatan narkoba [8] menyatakan bahwa, dengan menggunakan algoritma C4.5 untuk prediksi tersangka dapat memberikan informasi atau pengetahuan yang mudah dipahami karena aturan yang dihasilkan ditampilkan dalam bentuk pohon keputusan dan mudah untuk membaca aturan susunannya. Dari hasil penelitiannya, alamat tersangka kasus narkoba menjadi faktor yang paling berpengaruh terhadap tindak pidana narkoba. Bukti empiris menunjukkan bahwa, penelitiannya dilakukan menggunakan metode yang baik sehingga menghasilkan pengetahuan atau informasi *rule* yang mudah untuk dipahami. Namun, metode yang digunakan masih memiliki kelemahan sebagai berikut: Informasi yang dihasilkan tidak menunjukkan nilai akurasi yang jelas, sehingga dapat menimbulkan persepsi kurang baik terhadap tingkat kepercayaan publik pada penelitiannya. Selain itu, pemilihan

algoritma juga menjadi faktor penting dalam penelitian ini agar dapat meningkatkan kualitas prediksi terhadap sekumpulan data. Untuk menentukan algoritma apa yang paling tepat, dapat dilakukan perbandingan nilai performa terhadap akurasi prediksi dari berbagai algoritma yang diuji. Dari permasalahan ini, penulis mengusulkan metode komparasi 3 algoritma *decision Tree* (ID3, C4.5, CHAID) dengan uji validasi menggunakan *10-fold cross validation*.

A. ID3 dan C4.5

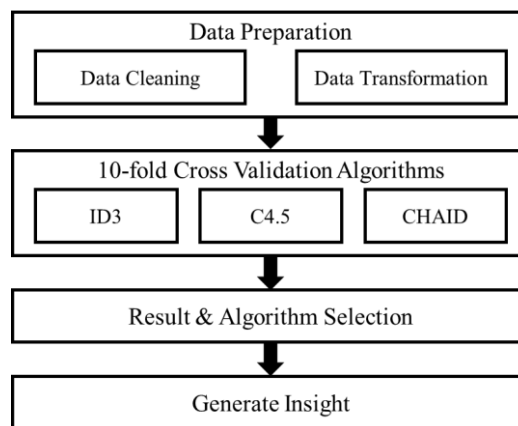
Iterative Dichotomiser 3 (ID3) sebagai algoritma pengklasifikasian secara hirarki memiliki struktur dasar yang berulang, dan *subset* yang dihasilkan dari *data training* akan dipilih secara acak untuk membentuk sebuah pohon keputusan [9]. Algoritma ini diusulkan oleh Quinlan dan dipublikasikan pada tahun 1986. Pohon keputusan ID3 sangat dipengaruhi oleh informasi *gain* yang didapatkan dari data yang diuji. Dalam upaya mendapatkan hasil yang maksimal, banyak peneliti melakukan *improvement* pada algoritma ini [10]. Dalam memprediksi daftar tersangka tindak kejahatan, ID3 dapat secara efektif dalam menganalisis data dari TKP dengan data yang sudah ada pada *dataset* [11]. Tidak sampai di sini, kemudian Quinlan memperkenalkan algoritma C4.5 dan dipublikasikan pada tahun 1993 [12] sebagai lanjutan dari algoritma ID3. Mulai saat itu, algoritma C4.5 menjadi acuan untuk model standar klasifikasi *supervised*. Sebagai perbaikan dari ID3, algoritma C4.5 menghasilkan pohon keputusan yang lebih luas dan efisien. Keunggulan ini bersifat relatif [13], tergantung *dataset* yang kita miliki sebagai penelitian.

B. CHAID

Teknik pohon keputusan *Chi-square Automatic Interaction Detection* (CHAID) yang merupakan teknik dengan menggunakan pengujian *bonferroni* untuk menyesuaikan uji signifikansi terhadap atribut data. Gordon V.Kass mengusulkan teknik ini, kemudian pada tahun 1980 telah dipublikasikan [14]. Dalam penerapannya, data akan dipartisi dan dipilih lalu dibandingkan untuk menemukan predictor terbaik dalam menganalisis data.

III. METODE PENELITIAN

Penelitian ini menggunakan metode perbandingan algoritma, dengan 3 algoritma *decision tree* yang divalidasi menggunakan *10-fold cross validation* sebagai perbandingan untuk mengetahui masing-masing performanya. Algoritma yang diuji performa akurasinya yaitu ID3, C4.5, dan CHAID. Dengan demikian, pemilihan algoritma untuk *profiling* tersangka kasus narkoba menjadi lebih tepat. Kerangka metode penelitian bisa dilihat pada gambar 1.



Gambar 1 Kerangka Metode Penelitian.

Sebagai catatan bahwa metode ini bukanlah sebuah model untuk memecahkan permasalahan pada analisis forensik secara keseluruhan, namun digunakan dalam proses analisis *profiling* pelaku penyalahgunaan narkoba di Jawa Barat. Data yang diuji merupakan data *nominal* untuk keseluruhan atribut pada *dataset*, walaupun pada tabel 1 terlihat ada atribut *numeric*, tetapi data tersebut bersifat *labeling* pada tingkat pendidikan yang tidak bisa diolah dengan operasi bilangan. Oleh karena itu, metode ini tidak bisa diterapkan pada *dataset* yang memiliki atribut *numeric* sebagai perhitungan, seperti algoritma ID3 dan CHAID.

IV. HASIL DAN PEMBAHASAN

A. Data Preparation

Tidak semua data yang dikumpulkan sebagai *dataset* sudah terstruktur dengan baik dan siap untuk dianalisis, kualitas hasil prediksi juga sangat tergantung pada kualitas data yang dimiliki. Maka dari itu proses *preparation* sangat penting sebelum melakukan pemodelan menggunakan algoritma. Data yang tersedia sebagai *data sampling* pada penelitian ini, merupakan data *private* tersangka tindak kejahatan di wilayah Jawa Barat dari tahun 2016 sampai dengan 2019. Kemudian proses *cleaning* dilakukan pada bagian ini, pemilihan atribut (*attributes selection*) dan transformasi data adalah cara yang diimplementasikan peneliti guna mempersiapkan data yang bersih dan siap untuk dilakukan pemodelan. Pada tabel 1 adalah atribut-atribut yang telah dipilih dalam penelitian ini.

Tabel 1 Atribut Terpilih

Atribut	Tipe Data
TKP	Nominal
Jenis Kelamin	Nominal
Usia	Nominal
Pendidikan	Numeric
Pekerjaan	Nominal
Kasus Narkoba	Nominal

Pada tabel 1 atribut “Pendidikan” memiliki tipe data *numeric*, karena pemodelan dengan algoritma ID3 dan CHAID tidak bisa menangani atribut *numeric*, maka proses transformasi data dilakukan pada tahap ini. Perubahan tipe data dari *numeric* ke *nominal* bisa dilihat pada tabel 2.

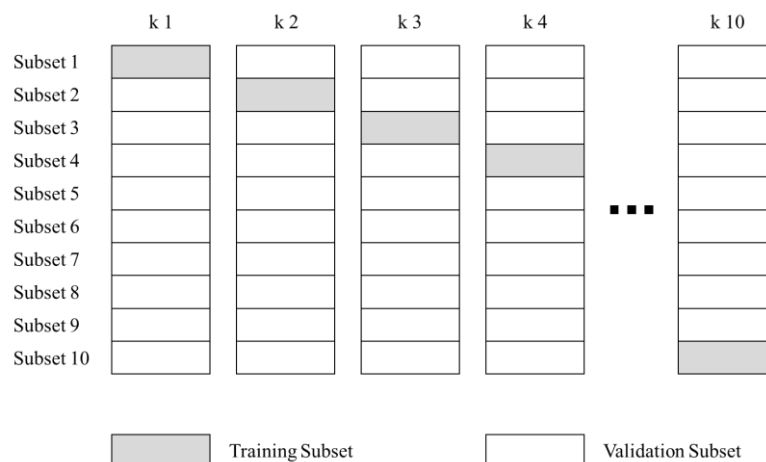
Tabel 2 Transformasi Data

Pendidikan	Pendidikan
0	TS
1	SD
2	SMP
3	SMA
4	DIPLOMA
5	S1
6	S2
7	S3

Terlihat pada tabel 2, telah dilakukan proses transformasi data, dalam prosesnya penulis menggunakan Microsoft Excel untuk memudahkan perubahan data tersebut.

B. *10-fold Cross Validation*

Dalam metode perbandingan, tentunya mengetahui performa dari masing-masing algoritma yang akan diuji terhadap *dataset* menjadi hal utama dalam penelitian ini. Tujuannya untuk mengetahui model terbaik yang akan digunakan dalam proses analisis sebuah data dengan melihat nilai akurasi dari masing-masing algoritma. Model algoritma yang akan diuji adalah 3 algoritma dari *decision tree* yaitu ID3, C.45, dan CHAID dengan menggunakan teknik *k-fold cross validation*. Dengan memisahkan *dataset* menjadi beberapa bagian dalam ukuran yang sama untuk membuat sampel data yang akan diuji berulang kali sesuai dengan jumlah *k* yang ditentukan. Penelitian ini menentukan nilai 10 untuk *k*, artinya *subset data training* dan validasi akan diuji sebanyak 10 kali secara acak terhadap model yang diterapkan pada algoritma (ID3,C4.5,CHAID). Proses pemodelan dan validasi, peneliti menggunakan RapidMiner versi 9.7. Pada gambar 2 menunjukkan ilustrasi dari *10-fold cross validation*.



Gambar 2 Ilustrasi 10-fold cross validation

Tahap selanjutnya adalah perhitungan terhadap tabel confusion matrix dari hasil 10-fold cross validation, dengan ilustrasi seperti yang terlihat pada tabel 3:

Tabel 3 Ilustrasi *Confusion Matrix*

		<i>Actual Class</i>	
<i>Predicted Class</i>	N	TN	FN
	P	FP	TP

Keterangan:

N (*Negative*), *TN* (*True Negative*), *FN* (*False Negative*), *P* (*Positive*), *FP* (*False Positive*), *TP* (*True Positive*).

- Validasi Algoritma ID3

Dengan menggunakan *10-fold cross validation* menghasilkan *confusion matrix* ID3 terlihat pada tabel 4.

Tabel 4 *Confusion Matrix* ID3

	<i>True (Tidak)</i>	<i>True (Ya)</i>
Pred. Tidak	703	263
Pred. Ya	232	460

Algoritma ID3 memiliki prediksi Tidak dan menghasilkan *True* (Tidak): 703, *True* (Ya): 263. Artinya dalam memprediksi profil tersangka yang bukan kasus narkoba, ID3 mempunyai nilai akurasi:

$$Pred(N) = \frac{TN}{\sum Actual Class N}$$

$$Pred(N) = \frac{703}{703 + 263}$$

Maka hasilnya adalah 0.7277 atau memiliki akurasi 72.77% untuk prediksi “Tidak”.

Sedangkan untuk prediksi “Ya” menggunakan algoritma ID3 menghasilkan *True* (Tidak): 232, *True* (Ya): 460, sehingga nilai akurasi untuk prediksi “Ya” menggunakan ID3 yaitu:

$$Pred(P) = \frac{TP}{\sum Actual Class P}$$

$$Pred(P) = \frac{460}{232 + 460}$$

Hasil yang didapat adalah 0.6647 atau 66.47% akurasi dalam memprediksi “Ya” pada ID3. Nilai prediksi “Ya” dan “Tidak” sudah diketahui, maka untuk menghitung total akurasi secara keseluruhan dari model algoritma ID3 yaitu sebagai berikut:

$$Akurasi = \frac{TN + TP}{\sum Actual Class}$$

$$Akurasi = \frac{703 + 460}{703 + 263 + 232 + 460}$$

Nilai akurasi keseluruhan pada pemodelan algoritma ID3 adalah 0,7014 atau **70.14%**. Pada tabel 5 akan terlihat hasil perhitungan akurasi dari ID3.

Tabel 5 Total Akurasi ID3

<i>Pred (Tidak)</i>	<i>Pred (Ya)</i>	<i>Total Akurasi</i>
72.77%	66.47%	70.14%

- Validasi Algoritma C4.5

Confusion Matrix C4.5 dengan menggunakan *10-fold cross validation* dapat dilihat pada tabel 6.

Tabel 6 *Confusion Matrix* C4.5

	<i>True (Tidak)</i>	<i>True (Ya)</i>
Pred. Tidak	816	338
Pred. Ya	119	385

Algoritma C4.5 dalam memprediksi “Tidak” untuk tersangka kasus narkoba yaitu *True (Tidak)*: 816, *True (Ya)*: 338, maka nilai akurasi dalam prediksi “Tidak”:

$$Pred(N) = \frac{TN}{\sum Actual\ Class\ N}$$

$$Pred(N) = \frac{816}{816 + 338}$$

Nilai akurasi C4.5 dalam memprediksi “Tidak” adalah 0.7071 atau 70.71%. Selanjutnya, akurasi C4.5 dalam memprediksi “Ya” menghasilkan nilai *True (Tidak)*: 119, *True (Ya)* 385, maka akurasi dalam prediksi “Ya” adalah:

$$Pred(P) = \frac{TP}{\sum Actual\ Class\ P}$$

$$Pred(P) = \frac{385}{119 + 385}$$

Hasilnya adalah 0.7639 atau 76.39% untuk nilai akurasi prediksi “Ya” pada C4.5. Total nilai akurasi menggunakan algoritma C4.5 adalah:

$$Akurasi = \frac{TN + TP}{\sum Actual\ Class}$$

$$Akurasi = \frac{816 + 385}{816 + 338 + 119 + 385}$$

Hasilnya 0.7244 atau akurasi menggunakan model algoritma C4.5 yaitu **72.44%**. Tabel 7 menunjukkan nilai akurasi dari C4.5.

Tabel 7 Total Akurasi C4.5

<i>Pred (Tidak)</i>	<i>Pred (Ya)</i>	<i>Total Akurasi</i>
70,71%	76.39%	72.44%

- Validasi Algoritma CHAID

Confusion matrix CHAID dengan *10-fold cross validation* dapat dilihat pada tabel 8.

Tabel 8 *Confusion Matrix* CHAID

	<i>True (Tidak)</i>	<i>True (Ya)</i>
Pred. Tidak	788	286
Pred. Ya	147	437

Nilai akurasi CHAID dalam memprediksi Tidak, yaitu *True (Tidak)*: 788, *True (Ya)*: 286, perhitungan akurasi adalah:

$$Pred(N) = \frac{TN}{\sum Actual\ Class\ N}$$

$$Pred(N) = \frac{788}{788 + 286}$$

Dari hasil perhitungan mendapatkan nilai 0.7337 atau 73.37% akurasi dalam memprediksi “Tidak” pada algoritma CHAID. Akurasi prediksi “Ya” menggunakan CHAID dengan berdasarkan nilai pada *confusion matrix* adalah *True* (Tidak): 147, *True* (Ya): 437, perhitungan akurasi “Ya” adalah:

$$Pred (P) = \frac{TP}{\sum Actual Class P}$$

$$Pred (P) = \frac{437}{147 + 437}$$

Prediksi “Ya” pada algoritma CHAID memiliki nilai 0.7483 atau memiliki nilai akurasi 74.83%. Nilai akurasi secara keseluruhan dengan menggunakan CHAID adalah:

$$Akurasi = \frac{TN + TP}{\sum Actual Class}$$

$$Akurasi = \frac{788 + 437}{788 + 286 + 147 + 437}$$

Jadi total akurasi dari algoritma CHAID sebesar 0.7389 atau **73.89%**. Pada tabel 9 menunjukkan semua nilai akurasi dari CHAID.

Tabel 9 Total Akurasi CHAID

<i>Pred (Tidak)</i>	<i>Pred (Ya)</i>	<i>Total Akurasi</i>
73.37%	74.83%	73.89%

C. Result & Algorithm Selection

Berdasarkan hasil perhitungan akurasi dari 3 algoritma yang telah dilakukan menggunakan *10-fold cross validation*, maka dapat dilihat pada tabel 10 untuk rekapitulasi hasil perhitungan nilai akurasinya.

Tabel 10 Rekapitulasi Perbandingan Nilai Akurasi

	<i>Pred (Tidak)</i>	<i>Pred (Ya)</i>	<i>Total Akurasi</i>
ID3	72.77%	66.47%	70.14%
C4.5	70.71%	76.39%	72.44%
CHAID	73.37%	74.83%	73.89%

Pada tabel 10 menunjukkan bahwa algoritma CHAID memiliki total akurasi tertinggi yaitu 73.89%. Dengan demikian, pemodelan dengan menggunakan algoritma CHAID pada penelitian ini merupakan model terbaik untuk melakukan klasifikasi atau prediksi dalam profiling tersangka kasus narkoba.

D. Generate Insight

Model algoritma telah terpilih, yaitu algoritma CHAID. Pada tahap selanjutnya adalah melakukan pemodelan dengan menggunakan RapidMiner versi 9.7 untuk mendapatkan *insight* dan *rule profiling* tersangka kasus narkoba. Berikut adalah *rule* profil tersangka kasus narkoba di Jawa Barat.

```
TKP = Bandung
| Usia = Dewasa akhir: Tidak {Tidak=12, Ya=3}
| Usia = Dewasa awal: Tidak {Tidak=15, Ya=4}
| Usia = Lansia akhir: Tidak {Tidak=4, Ya=0}
| Usia = Lansia awal: Tidak {Tidak=6, Ya=1}
| Usia = Remaja akhir
| | Pendidikan = DIPLOMA
| | | Pekerjaan = Pelajar / mahasiswa: Tidak {Tidak=1, Ya=1}
| | | Pekerjaan = Wiraswasta: Ya {Tidak=1, Ya=3}
| | Pendidikan = SD: Tidak {Tidak=2, Ya=0}
| | Pendidikan = SMA: Tidak {Tidak=5, Ya=1}
| | Pendidikan = SMP: Tidak {Tidak=13, Ya=4}
TKP = Cimahi
| Pendidikan = DIPLOMA
| | Usia = Dewasa akhir: Ya {Tidak=2, Ya=12}
| | Usia = Dewasa awal: Ya {Tidak=10, Ya=21}
| | Usia = Lansia awal: Tidak {Tidak=4, Ya=2}
| | Usia = Remaja akhir: Ya {Tidak=1, Ya=18}
```

```
| Pendidikan = S1: Tidak {Tidak=3, Ya=3}
| Pendidikan = S2: Ya {Tidak=1, Ya=5}
| Pendidikan = SD: Tidak {Tidak=2, Ya=2}
| Pendidikan = SMA: Ya {Tidak=16, Ya=33}
| Pendidikan = SMP: Tidak {Tidak=17, Ya=9}
TKP = Cirebon: Tidak {Tidak=33, Ya=9}
TKP = Depok
| Pendidikan = DIPLOMA: Ya {Tidak=13, Ya=152}
| Pendidikan = S1: Ya {Tidak=1, Ya=9}
| Pendidikan = S2: Ya {Tidak=1, Ya=13}
| Pendidikan = SD
| | Jenis Kelamin = Laki-laki
| | | Usia = Dewasa awal: Ya {Tidak=3, Ya=6}
| | | Usia = Remaja akhir: Tidak {Tidak=3, Ya=1}
| | Jenis Kelamin = Perempuan: Tidak {Tidak=2, Ya=0}
| Pendidikan = SMA: Ya {Tidak=15, Ya=54}
| Pendidikan = SMP
| | Jenis Kelamin = Laki-laki: Ya {Tidak=9, Ya=13}
| | Jenis Kelamin = Perempuan: Tidak {Tidak=2, Ya=0}
TKP = Garut: Tidak {Tidak=70, Ya=9}
TKP = Kuningan
| Pendidikan = DIPLOMA: Ya {Tidak=17, Ya=39}
| Pendidikan = S1: Tidak {Tidak=2, Ya=1}
| Pendidikan = S2: Ya {Tidak=4, Ya=5}
| Pendidikan = SD: Tidak {Tidak=9, Ya=5}
| Pendidikan = SMA: Ya {Tidak=13, Ya=30}
| Pendidikan = SMP
| | Usia = Dewasa akhir: Ya {Tidak=5, Ya=5}
| | Usia = Dewasa awal: Ya {Tidak=6, Ya=8}
| | Usia = Lansia akhir: Tidak {Tidak=2, Ya=0}
| | Usia = Lansia awal: Tidak {Tidak=6, Ya=1}
| | Usia = Remaja akhir: Ya {Tidak=3, Ya=5}
TKP = Majalengka: Tidak {Tidak=27, Ya=7}
TKP = Metro Bekasi: Ya {Tidak=1, Ya=1}
TKP = Purwakarta
| Pendidikan = DIPLOMA: Ya {Tidak=0, Ya=7}
| Pendidikan = SD: Ya {Tidak=1, Ya=2}
| Pendidikan = SMA
| | Usia = Dewasa awal: Ya {Tidak=0, Ya=5}
| | Usia = Lansia akhir: Tidak {Tidak=2, Ya=0}
| | Usia = Remaja akhir: Ya {Tidak=0, Ya=4}
| Pendidikan = SMP: Ya {Tidak=1, Ya=6}
TKP = Sukabumi: Tidak {Tidak=391, Ya=127}
TKP = Tasikmalaya: Tidak {Tidak=178, Ya=77}
```

Rules yang dihasilkan menunjukkan bahwa TKP merupakan atribut *predictor* yang paling signifikan terhadap atribut target atau *label*, sehingga dijadikan *root* pertama dalam pohon keputusan pada algoritma CHAID. Tingkat pendidikan lalu diikuti usia, menjadi faktor kedua dan ketiga setelah TKP sebagai atribut *predictor* penting pada profil tersangka kasus narkoba di Jawa Barat.

KESIMPULAN

Dari hasil uji performa menggunakan *10-fold cross validation*, algoritma CHAID memiliki nilai akurasi tertinggi yaitu 73.89% sedangkan C4.5 akurasi 72.44%, lalu nilai akurasi paling rendah didapatkan oleh ID3 sebesar 70.14%. Hasil klasifikasi dengan algoritma CHAID sebagai pemodelan yang terpilih, menunjukkan bahwa: TKP, pendidikan, dan usia menjadi atribut yang paling memengaruhi tersangka kasus narkoba di Jawa Barat. Dari serangkaian analisis yang dilakukan, metode perbandingan tiga algoritma *decision tree* (ID3, C4.5, CHAID) menjadi solusi dalam menentukan pemilihan model algoritma yang tepat untuk menganalisis data terhadap *dataset* yang dimiliki. Akan tetapi penelitian ini memiliki kelemahan yaitu algoritma yang digunakan dalam metode komparasi belum bisa dilakukan terhadap data yang memiliki atribut *numeric*. Penelitian selanjutnya diharapkan dapat menerapkan metode ini dengan menambahkan metode transformasi data yang tepat, sehingga dapat menangani data *numeric* yang dapat meningkatkan kualitas prediksi.

DAFTAR PUSTAKA

- [1] F. C. Tsai, M. C. Hsu, C. T. Chen, dan D. Y. Kao, "Exploring drug-related crimes with social network analysis," *Procedia Comput. Sci.*, vol. 159, hal. 1907–1917, 2019.

- [2] PUSLITDATIN, "Penggunaan Narkotika di Kalangan Remaja Meningkat," *BNN*, 2019.
- [3] Z. Qiao, T. Chai, Q. Zhang, X. Zhou, dan Z. Chu, "Predicting potential drug abusers using machine learning techniques," *ICIIBMS 2019 - 4th Int. Conf. Intell. Informatics Biomed. Sci.*, hal. 283–286, 2019.
- [4] L. McClendon dan N. Meghanathan, "Using Machine Learning Algorithms to Analyze Crime Data," *Mach. Learn. Appl. An Int. J.*, vol. 2, no. 1, hal. 1–12, 2015.
- [5] N. H. M. Shamsuddin, N. A. Ali, dan R. Alwee, "An overview on crime prediction methods," *6th ICT Int. Student Proj. Conf. Elev. Community Through ICT, ICT-ISPC 2017*, vol. 2017-Janua, no. May, hal. 1–5, 2017.
- [6] L. Jayanti, S. R. Sentinuwo, O. A. Lantang, dan A. Jacobus, "Analisa Pola Penyalahgunaan Facebook Sebagai Alat Kejahatan Trafficking Menggunakan Data Mining," *J. Tek. Inform.*, vol. 8, no. 1, 2016.
- [7] D. Gultom *dkk.*, "Penerapan Algoritma K-Means Untuk Mengetahui Tingkat Tindak Kejahatan Daerah Pematangsiantar," *J. Teknol. Inf.*, vol. 4, no. 1, hal. 146–151, 2020.
- [8] S. Wahyuni, "Implementation of Data Mining to Analyze Drug Cases Using C4.5 Decision Tree," *J. Phys. Conf. Ser.*, vol. 970, no. 1, hal. 0–6, 2018.
- [9] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, hal. 81–106, 1986.
- [10] I. D. Mienye, Y. Sun, dan Z. Wang, "Prediction performance of improved decision tree-based algorithms: A review," *Procedia Manuf.*, vol. 35, hal. 698–703, 2019.
- [11] A. O. Ogunde, G. O. Ogunleye, dan O. Oreoluwa, "A Decision Tree Algorithm Based System for Predicting Crime in the University," *Mach. Learn. Res.*, vol. 2, no. 1, hal. 26–34, 2017.
- [12] J. R. Quinlan, "Programs for machine learning," *Inf. Control*, no. 1, hal. 235–240, 1993.
- [13] H. Hao, T. Chen, J. Lu, J. Liu, dan X. Ma, "The Research and Analysis in Decision Tree Algorithm Based on C4.5 Algorithm," *Proc. 2018 IEEE 3rd Adv. Inf. Technol. Electron. Autom. Control Conf. IAEAC 2018*, no. 2, hal. 1882–1886, 2018.
- [14] G. V. Kass, "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Appl. Stat.*, vol. 29, no. 2, hal. 119, 1980.