

## Penerapan dan Perbandingan Tiga Metode Analisis Pohon Keputusan pada Klasifikasi Penderita Kanker Payudara

Jody Alwin Irawadi\*, Siti Sunendiari

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan, Universitas Islam Bandung, Indonesia.

\*jodyalwin20@gmail.com, diarisunen22@gmail.com

**Abstract.** Today there is a considerable amount of work dealing with decision trees, especially in survival analysis (Ibrahim et al. 2008). Cases classified as survival analysis, like cancer patients. This study discusses the application of data mining which is to obtain diagnostic results. The classification technique uses information obtained from medical records of breast cancer patients in Yugoslavia. A method for answering these problems through decision tree analysis using the CHAID, Exhaustive CHAID and CART methods. Empirically aiming to compare performance of three decision tree classification methods so that the best method is obtained. It was concluded that best method used in applying to the classification of breast cancer sufferers was the CART method because it was able to get the most significant variables at most four, namely inv-node, tumor size, deg-malig and breast parts. Then it has a total accuracy rate with highest value of 84.9 percent and has a total error rate with lowest value of 15.1 percent.

**Keywords:** Exhaustive CHAID, CART, breast cancer.

**Abstrak.** Dewasa ini ada cukup banyak pekerjaan yang berurusan dengan pohon keputusan, terutama dalam analisis survival (Ibrahim dkk, 2008). Kasus yang tergolong analisis survival seperti penderita penyakit kanker. Penelitian ini membahas mengenai penerapan data mining yang digunakan untuk mendapatkan hasil diagnostik. Pendekatan teknik klasifikasi dengan menggunakan informasi yang diperoleh pada rekam medis data penderita kanker payudara di Yugoslavia. Salah satu metode untuk menjawab permasalahan tersebut melalui analisis pohon keputusan dengan metode CHAID, Exhaustive CHAID dan CART. Secara empiris bertujuan untuk membandingkan kinerja tiga metode pengklasifikasi pohon keputusan agar didapatkan metode manakah yang terbaik. Maka disimpulkan bahwa metode terbaik yang digunakan dalam penerapan pada klasifikasi penderita kanker payudara adalah metode CART sebab mampu mendapatkan variabel signifikan yang paling banyak ada empat, yakni inv-node, ukuran tumor, deg-malig dan bagian payudara. Kemudian memiliki tingkat akurasi total dengan nilai tertinggi sebesar 84.9 persen dan memiliki total tingkat kesalahan dengan nilai yang terendah sebesar 15.1 persen.

**Kata Kunci:** Exhaustive CHAID, CART, kanker payudara.

## A. Pendahuluan

Aktifitas penelitian kanker pada umumnya bersifat klinis atau biologis, statistik yang didorong oleh data penelitian menjadi sebuah pelengkap (Delen dkk, 2004). Menurut Globocan yang dirilis *International Agency for Research on Cancer (IARC)* menyebutkan diperkirakan bahwa jumlah kasus insiden kanker di tahun 2040 mencapai 29,5 juta kasus. Pada tahun 2018, *World Health Organization (WHO)* menyatakan bahwa kanker sebagai penyebab utama kematian kedua di dunia. Berdasarkan data Depkes tahun 2019, angka kejadian penyakit kanker di Indonesia 136,2 per 100.000 penduduk dan menempati urutan ke 23 di Asia.

Salah satu metode statistika yang dapat digunakan untuk menjawab permasalahan ini adalah data mining dengan analisis pohon keputusan (*decision tree*), sebab banyak digunakan dalam bidang yang membutuhkan tanggung jawab tinggi, seperti pada diagnostik kedokteran dan analisa keuangan. Dewasa ini ada cukup banyak pekerjaan yang berurusan dengan pohon keputusan, terutama dalam analisis *survival* (Ibrahim dkk, 2008). Kasus yang tergolong analisis *survival* yakni pada penderita penyakit kanker.

Tiga metode yang digunakan dalam penelitian ini *Chi-square Automatic Interaction Detection (CHAID)*, *Exhaustive CHAID* dan *Classification and Regression Tree (CART)*. CHAID menggunakan uji *Chi-square* antara peubah respon dengan penjelas yang umumnya data kategorik dan jika terdapat penggabungan pada peubah penjelas maka dilakukan koreksi *Bonferroni*, sementara CART terdiri dari dua analisis yaitu menghasilkan pohon klasifikasi jika variabel respon bertipe kategorik sedangkan jika variabel respon bertipe kontinu/numerik maka akan menghasilkan pohon regresi.

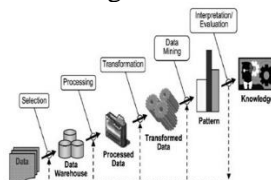
Salah satu metode diatas yakni CHAID, mengalami pengembangan metode yang dinamakan *Exhaustive CHAID* yang diharapkan mampu mengatasi peubah penjelas yang memiliki banyak kategori dalam menghadapi peubah penjelas yang mengambang. Sehingga dalam penelitian ini akan menggunakan tiga metode analisis pohon keputusan yakni CHAID, *Exhaustive CHAID* dan CART. Secara empiris bertujuan membandingkan kinerja tiga metode pengklasifikasi pohon keputusan agar didapatkan metode manakah yang terbaik untuk mendapatkan hasil diagnostik pada kasus penderita kanker payudara.

Berdasarkan uraian diatas, tujuan yang ingin dicapai dari penelitian ini adalah:

1. Mengetahui faktor-faktor apa saja yang mempengaruhi penderita kanker payudara berdasarkan hasil analisis pohon keputusan menggunakan metode CHAID, *Exhaustive CHAID* dan CART.
2. Mengetahui hasil klasifikasi penderita kanker payudara berdasarkan analisis pohon keputusan menggunakan metode CHAID, *Exhaustive CHAID* dan CART.
3. Membandingkan kinerja tiga metode pengklasifikasi penderita kanker payudara berdasarkan analisis pohon keputusan menggunakan metode CHAID, *Exhaustive CHAID* dan CART.

## B. Landasan Teori

Menurut Larose (2005), data mining adalah cara mengekstraksi informasi yang implisit, baru dan potensial dari data. Dengan cara mengeksplorasi dan menganalisis data untuk menemukan pola penting yang terkandung didalamnya. Tahapan yang dilakukan pada proses data mining menurut Fayyad (1996) pada Gambar 1 sebagai berikut:



**Gambar 1.** Tahapan Proses Data Mining

Dengan catatan pada tahap data mining nantinya sekumpulan baris data akan dibuat data *training* dan data *testing*. Data *training* untuk membuat model sedangkan data *testing* untuk memvalidasi model yang dibuat. Menurut Venkatesan dan Velmurugan (2015) dalam

penelitiannya bahwa data *training* menggunakan 2/3 dari baris data dan 1/3 dari baris data untuk data *testing*.

Pohon keputusan atau *decision tree* menurut Cho dan Ngai (2003) merupakan suatu kondisi yang digunakan untuk mengklasifikasi objek yang biasa ditampilkan melalui gambar daun dan cabang. Daun (*leaf*) pada pohon keputusan diidentifikasi oleh kelas, sementara cabangnya (*internal node*) mewakili suatu kondisi dari atribut objek yang terukur. Level *node* teratas dari sebuah pohon keputusan adalah akar (*root*) yang biasanya berupa atribut yang paling memiliki pengaruh terbesar pada suatu kelas tertentu. Pohon keputusan merupakan salah satu teknik klasifikasi yang paling populer dalam data mining. Pohon keputusan sendiri memiliki beberapa metode diantaranya CHAID, *Exhaustive* CHAID, dan CART yang akan dibahas dalam penelitian ini.

Metode *Chi-Squared Automatic Interaction Detection* (CHAID) digunakan untuk menyelidiki struktur keterkaitan antar peubah, yaitu antar peubah respon kategorik dengan peubah penjelas yang umumnya kategorik (Alamundi, 1998). CHAID termasuk dalam analisis statistika non parametrik. Metode CHAID menggunakan kriteria chi kuadrat dalam pengoperasiannya (Kass, 1980). Uji chi kuadrat digunakan untuk mengetahui independensi, yaitu apakah terdapat hubungan antara dua variabel tertentu atau tidak pada tiap levelnya.

Menurut Kass (1980), tahapan yang dilakukan pada metode CHAID terbagi tiga tahap:

1. Tahap penggabungan, pada tahap ini dilakukan pemeriksaan terhadap nilai chi kuadrat peubah respon terhadap masing-masing peubah penjelas. Langkah pertama adalah membuat tabulasi silang antara peubah respon dengan masing-masing peubah penjelas. Langkah kedua memasangkan peubah-peubah penjelas dengan peubah respon yang berhubungan nyata. Langkah berikutnya adalah menghitung chi kuadrat ( $\chi^2$ ) tiap pasangan kategori dengan rumus (Agresti, 1990):

$$\chi^2 = \sum_{i=1}^b \sum_{j=1}^k \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Setelah menghitung chi kuadrat, melakukan uji hipotes pada setiap pasangan kategori. Hipotesis pada uji chi kuadrat adalah:

$H_0$  : peubah respon dan peubah penjelas saling bebas.

$H_1$  : peubah respon dan peubah penjelas tidak saling bebas.

Keputusan yang diambil pada uji hipotesis ini adalah tolak  $H_0$  jika nilai  $\chi^2$  hitung  $> \chi^2_{(\alpha; (r-1)(c-1))}$ . Langkah berikutnya yaitu menghitung nilai p untuk setiap pasangan. Apabila ada pasangan kategorik yang memiliki nilai  $\chi^2$  hitung  $< \chi^2_{(\alpha; (r-1)(c-1))}$  maka dilakukan penggabungan pasangan dengan nilai chi kuadrat terkecil dan nilai p terbesar kedalam satu kategori baru. Langkah terakhir dari tahap ini adalah memeriksa kembali nilai chi kuadrat kategori baru setelah digabung dengan kategori lainnya jika masih terdapat pasangan yang nilai  $\chi^2$  hitung  $< \chi^2_{(\alpha; (r-1)(c-1))}$  maka kembalikan ke langkah sebelumnya.

2. Tahap pemisahan merupakan tahap memilih peubah yang digunakan sebagai pemisah simpul terbaik. Peubah penjelas yang memiliki nilai p-value terkecil menjadi peubah yang digunakan sebagai pemisah simpul terbaik.
3. Tahap penghentian dilakukan jika sudah tidak ada hubungan nyata antara peubah penjelas dengan peubah respon dan ukuran dari anak simpul kurang dari nilai ukuran anak simpul minimum spesifikasi. Jika terdapat penggabungan pada peubah penjelas maka dilakukan koreksi *Bonferroni* untuk masing-masing jenis peubah bebas adalah sebagai berikut:

- Pengali *Bonferroni* tipe bebas untuk peubah asal berskala nominal.

$$B = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{i!(r-i)!}$$

- Pengali *Bonferroni* tipe monotonik untuk peubah asal berskala ordinal.

$$B = \binom{c-1}{r-1}$$

- Pengali *Bonferroni* tipe mengambang untuk peubah asal berskala ordinal namun terdapat kategori yang belum dapat ditentukan urutannya.

$$B = \binom{c-2}{r-2} + r \binom{c-2}{r-1}$$

Metode CHAID mengalami perkembangan algoritma, untuk penyesuaian *Bonferroni* pada jumlah kategori lebih besar dari empat dan peubah penjas mengambang (Zhang dkk, 2014). *Exhaustive* CHAID dilakukan dengan mereduksi  $c$  kecil kategori menjadi  $c - 1$  kategori dengan cara menggabungkan kategori yang paling tidak nyata. Pada peubah penjas mengambang, kategori mengambang dikeluarkan sementara selama proses reduksi kategori. Selanjutnya digabung dengan salah satu kategori serupa, bila tidak serupa maka kategori mengambang dikeluarkan (Soemartojo, 2002). Pengali *Bonferroni* pada *Exhaustive* CHAID adalah:

$$N_{BC}(c) = 1 + \sum_{r=1}^{c-2} (c - r + 1, c - r)$$

$N(c - r + 1, c - r) = c - r$  untuk skala ordinal.

$$N(c - r + 1, c - r) = \sum_{v=0}^{c-r-1} (-1)^v \frac{(c-r-v)^{c-r+1}}{v!(c-r-v)!}$$
 untuk skala nominal.

$N(c - r + 1, c - r)$  adalah banyaknya cara untuk mengelompokkan  $(c - r + 1)$  kategori menjadi  $(c - r)$  kategori. Pada peubah mengambang, pengali *Bonferroni* yang digunakan adalah pengali *Bonferroni* pada peubah ordinal, tetapi ketika peubah mengambang digabung penyesuaian koefisien *Bonferroni* nya adalah:

$$N_{BC}(c) = N_{BC}(c - 1) + (c - 1) + 1$$

*Classification And Regresion Trees* (CART) adalah metode statistik yang digunakan untuk melakukan analisis klasifikasi. Tujuan dari CART adalah untuk mengklasifikasikan obyek menjadi dua atau lebih kelompok. Bagi sekumpulan data yang terdiri dari  $p$  buah variabel bebas dan sebuah variabel tak bebas, maka ketika yang dimiliki bertipe kategorik CART menghasilkan pohon klasifikasi (*classification trees*), sedangkan jika variabel tak bebas yang dimiliki bertipe kontinu atau numerik maka CART menghasilkan pohon regresi (*regression trees*).

Menurut Sumartini dan Purnami (2015), metode CART melalui tiga tahapan, yaitu pembentukan pohon klasifikasi, pemangkasan pohon klasifikasi dan penentuan pohon klasifikasi optimum:

1. Pembentukan pohon klasifikasi

Tahap ini diawali dengan menentukan variabel dan *threshold* untuk dijadikan pemilah tiap simpul. Tahapan pembentukan pohon klasifikasi terdiri dari:

- Pemilahan Pemilah

Fungsi Indeks Gini ditunjukkan pada persamaan:

$$i(t) = \sum_{i,j=1} p(j|t)p(i|t), i \neq j$$

Tahapan berikutnya adalah menentukan kriteria *goodness of split* untuk mengevaluasi pemilah dari pemilah  $s$  pada simpul  $t$  dengan rumus:

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

Pemilah yang menghasilkan  $\emptyset(s, t)$  lebih tinggi merupakan pemilah terbaik karena mampu mereduksi heterogenitas lebih tinggi.

- Penentuan Simpul Terminal

Pengembangan pohon akan berhenti apabila pada simpul terdapat pengamatan berjumlah kurang dari atau sama dengan 5 ( $n \leq 5$ ).

- Penandaan Label Kelas

Penentuan label kelas pada simpul terminal berdasarkan aturan jumlah terbanyak, yaitu jika

$$p(j_0|t) = \max_j \frac{N_i(t)}{N(t)}$$

2. Pemangkasan pohon klasifikasi

Ukuran pohon yang layak dapat dilakukan dengan pemangkasan pohon dengan ukuran *cost complexity minimum*.

$$R_\alpha(T) = R(T) + \alpha |\tilde{T}|$$

3. Penentuan pohon klasifikasi optimum

Penduga pengganti yang sering digunakan apabila pengamatan yang ada tidak cukup besar adalah *Cross Validation Estimate*. Pohon  $T^{(v)}$  dibentuk dari data training ke- $v$  dengan  $v=1,2,\dots,V$ . Dimisalkan  $d^{(v)}(x)$  adalah hasil pengklasifikasian, maka penduga sampel uji untuk  $R(T_t^{(v)})$  adalah sebagai berikut:

$$R(T_t^{(v)}) = \frac{1}{Nv} \sum_{(x_n, j_n) \in L_V} X(d^{(v)}(x_n) \neq j_n)$$

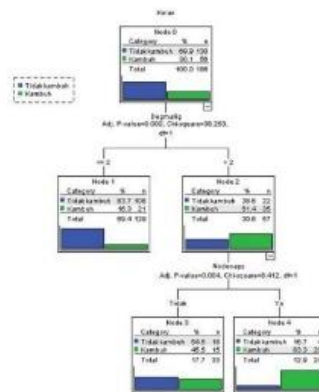
Menurut Yunindya dkk (2017), pohon klasifikasi optimal yang telah terbentuk dilakukan evaluasi dari hasil klasifikasi. Cara untuk mengevaluasi hasil klasifikasi adalah dengan menghitung akurasi klasifikasi *sensitivity* menggambarkan akurasi pada sampel kelas  $i$ , sedangkan *specificity* menggambarkan bagaimana akurasi pada sampel kelas  $j$ .

**C. Hasil Penelitian dan Pembahasan**

Data yang digunakan dalam penelitian ini adalah data sekunder, yang disediakan oleh Matjaz Zwitter dan Milan Soklic, Pusat Medis Universitas, Institut Onkologi, Ljubljana, Yugoslavia tahun 1988. Data yang diperoleh berupa data rekam medis mengenai penderita kanker payudara dan dipublikasikan oleh website *repository ics.uci.edu*. Keseluruhan variabel bertipe kategorik dengan variabel penjelas meliputi usia, status menopause, ukuran tumor, *inv-node*, *node-caps*, *deg-malig*, bagian payudara, kuadran payudara dan iradiasi, sedangkan variabel respon berupa kelas penderita.

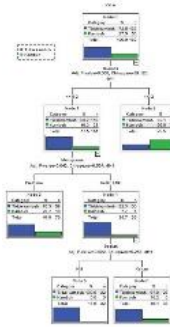
Data mengenai penderita kanker payudara ini ada sebanyak 286 data. Dalam beberapa variabel didapati adanya data *missing*, maka cara penanganannya dengan membuang data tersebut. Sehingga data yang dipakai sebanyak 277 data, kemudian dilakukan pengkodean. Langkah selanjutnya adalah membagi dua kelompok yaitu data *training* sebesar 2/3 dan data *testing* sebesar 1/3. Pembagian data dilakukan secara acak dengan bantuan *software*.

Pembagian dua kelompok data pada metode CHAID didapati 186 data *training* dan 91 data *testing*. Selanjutnya melakukan prosedur analisis. Hasil pembentukan pohon keputusan menggunakan metode CHAID disajikan pada Gambar 2. Terdapat lima node, dengan jumlah terminal node sebanyak tiga. Kedalaman dari pohon klasifikasi tersebut adalah dua. Diketahui variabel yang signifikan terhadap kelas penderita kanker payudara adalah *deg-malig* dan *node-caps*.



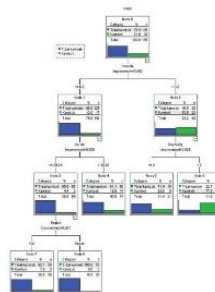
**Gambar 2.** Diagram Pohon Metode CHAID

Pembagian dua kelompok data pada metode *Exhaustive* CHAID didapati 182 data *training* dan 95 data *testing*. Selanjutnya melakukan prosedur analisis. Hasil pembentukan pohon keputusan menggunakan metode *Exhaustive* CHAID disajikan pada Gambar 3. Terdapat tujuh node, dengan jumlah terminal node sebanyak empat. Kedalaman dari pohon klasifikasi tersebut adalah tiga. Diketahui variabel yang signifikan terhadap kelas penderita kanker payudara adalah *inv-node*, menopause dan bagian payudara.



**Gambar 3.** Diagram Pohon Metode *Exhaustive* CHAID

Pembagian dua kelompok data pada metode CART didapati 185 data *training* dan 92 data *testing*. Selanjutnya melakukan prosedur analisis. Hasil pembentukan pohon keputusan menggunakan metode CART disajikan pada Gambar 4. Terdapat sembilan node, dengan jumlah terminal node sebanyak lima. Kedalaman dari pohon klasifikasi tersebut adalah tiga. Diketahui variabel yang signifikan terhadap kelas penderita kanker payudara adalah *inv-node*, ukuran tumor, *deg-malig* dan bagian payudara.



**Gambar 4.** Diagram Pohon Metode CART

Menghitung ketepatan klasifikasi pada masing-masing metode terhadap data *training* maupun data *testing*. Disajikan pada Tabel 1, 2 dan 3.

**Tabel 1.** Ketepatan Klasifikasi Pada Metode CHAID

		Data Training		Data Testing	
		Prediksi		Prediksi	
		Tidak kambuh	Kambuh	Tidak kambuh	Kambuh
Observasi	Tidak kambuh	126	4	63	3
	Kambuh	36	20	22	3
Tingkat Akurasi Total		0.785		0.725	
Total Tingkat Kesalahan		0.215		0.275	
<i>Sensitivity</i> (Se)		0.969		0.954	
<i>Specificity</i> (Sp)		0.357		0.120	

**Tabel 2.** Ketepatan Klasifikasi Pada Metode *Exhaustive* CHAID

		Data Training		Data Testing	
		Prediksi		Prediksi	
		Tidak kambuh	Kambuh	Tidak kambuh	Kambuh
Observasi	Tidak kambuh	118	14	48	16
	Kambuh	23	27	20	11
Tingkat Akurasi Total		0.797		0.621	
Total Tingkat Kesalahan		0.203		0.379	
<i>Sensitivity</i> (Se)		0.893		0.750	
<i>Specificity</i> (Sp)		0.540		0.355	

**Tabel 3.** Ketepatan Klasifikasi Pada Metode CART

		Data Training		Data Testing	
		Prediksi		Prediksi	
		Tidak kambuh	Kambuh	Tidak kambuh	Kambuh
Observasi	Tidak kambuh	140	5	48	3
	Kambuh	23	17	31	10
Tingkat Akurasi Total		0.849		0.630	
Total Tingkat Kesalahan		0.151		0.370	
<i>Sensitivity</i> (Se)		0.965		0.941	
<i>Specificity</i> (Sp)		0.425		0.244	

Melalui Gambar 2, 3 dan 4, terlihat diagram pohon yang sudah terbentuk pada masing-masing metode. Dapat diketahui variabel yang signifikan terhadap kelas penderita kanker payudara pada metode CHAID ada sebanyak dua variabel, pada metode *Exhaustive* CHAID ada sebanyak tiga variabel dan pada metode CART ada sebanyak empat variabel.

Memerhatikan data *training* pada Tabel 1, 2 dan 3 dengan melihat masing-masing metode dapat diketahui bahwa pohon klasifikasi yang terbentuk dengan tepat pengamatan yang tertinggi dimiliki oleh metode CART dengan nilai 84.9 persen, kemudian metode CART juga memiliki total tingkat kesalahan yang terendah dengan 15.1 persen. Sedangkan tingkat akurasi tertinggi pada penderita kanker payudara yang tidak kambuh (*sensitivity*) dimiliki oleh metode CHAID dengan 96.9 persen dan tingkat akurasi tertinggi pada penderita kanker payudara yang kambuh (*specificity*) dimiliki oleh metode *Exhaustive* CHAID dengan 54 persen.

Maka dari itu dapat ditentukan bahwa metode terbaik diantara tiga metode analisis pohon keputusan ialah metode CART sebab mampu mendapatkan variabel signifikan yang paling banyak, kemudian memiliki tingkat akurasi total dengan nilai tertinggi dan memiliki total tingkat kesalahan dengan nilai yang terendah dibanding dengan metode lainnya.

#### D. Kesimpulan

Berdasarkan pembahasan dalam penelitian ini, diperoleh kesimpulan sebagai berikut:

1. Berdasarkan diagram pohon masing-masing metode, bahwa faktor-faktor apa saja yang mempengaruhi penderita kanker payudara sebagaimana pada metode CHAID variabel yang signifikan terhadap kelas penderita kanker payudara ada dua variabel, yakni *deg-malig* dan *node-caps*. Pada metode *Exhaustive* CHAID variabel yang signifikan terhadap kelas penderita kanker payudara ada tiga variabel, yakni *inv-node*, status menopause dan bagian payudara. Kemudian pada metode CART variabel yang signifikan terhadap kelas penderita kanker payudara ada empat, yakni *inv-node*, ukuran tumor, *deg-malig* dan bagian payudara.
2. Berdasarkan ketepatan klasifikasi, pada masing-masing metode didapati metode CHAID

memprediksi dengan tepat pengamatan sebesar 78.5 persen. Pada metode Exhaustive CHAID memprediksi dengan tepat pengamatan sebesar 79.7 persen. Kemudian pada metode CART memprediksi dengan tepat pengamatan sebesar 84.9 persen.

3. Perbandingan diantara tiga metode analisis pohon keputusan memutuskan metode terbaik yang digunakan dalam penerapan pada klasifikasi penderita kanker payudara adalah metode CART sebab mampu mendapatkan variabel signifikan yang paling banyak, kemudian memiliki tingkat akurasi total dengan nilai tertinggi dan memiliki total tingkat kesalahan dengan nilai yang terendah dibanding dengan metode lainnya.

### Daftar Pustaka

- [1] Agresti, A. 1990. *Categorical Data Analysis*. USA: University of Florida.
- [2] Alamundi, A., Wigena, A. H. & Aunuddin. 1998. *Eksplorasi Struktur Data Menggunakan Metode CHAID*. Forum Statistika dan Komputasi. Institut Pertanian Bogor. ISSN: 08538115.
- [3] Breiman, L., Friedman, J., Olshen, R. & Stone, C. 2009. *Classification and Regression Trees*. New York – London: Chapman Hall.
- [4] Cho, V. & Ngai, E. 2003. *Data Mining for Selection of Insurance Sales Agents*. Expert Systems, 20, 123-132, doi: 10.1111/1468-0394.00235.
- [5] Delen, D., Walker, G. & Kadam, A. 2004. *Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods*. Artificial Intelligence in Medicine, 1-15, doi: 10.1016/j.artmed.2004.07.002.
- [6] Departemen Kesehatan. 2019. *Hari Kanker Sedunia 2019* (Online), (<https://depkes.go.id/article/view/19020100003/hari-kanker-sedunia-2019.html>, diakses 20 Oktober 2019).
- [7] Fayyad, U. 1996. *Advances in Knowledge Discovery and Data Mining*. MIT Press.
- [8] Globocan. 2018. *Cancer Tomorrow* (Online), (<https://gco.iarc.fr/>, diakses 20 Oktober 2019).
- [9] Halodoc. 2019. *Kanker Payudara* (Online), (<https://halodoc.com/kesehatan/kanker-payudara> diakses 30 Oktober 2019).
- [10] Ibrahim, N. A., Kudus, A., Daud, I. & Abu Bakar, M. R. 2008. *Decision Tree for Competing Risks Survival Probability in Breast Cancer Study*. International Journal of Biomedical Sciences, 3(1):25-29.
- [11] Kass, G. V. 1980. *An Exploratory Technique For Large Quantities For Categorical Analysis Data*. App Statis. 29, No 2 : 119 – 127.
- [12] Larose, D. T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Willey & Sons, Inc.
- [13] Nazar, R. R. 2018. *Penerapan Metode CHAID (Chi-Squared Automatic Interaction Detection) dan CART (Classification And Regression Trees) Pada Klasifikasi Preeklampsia*. Skripsi S1 Program Studi Statistika, Universitas Islam Indonesia.
- [14] Nugraha, J. 2016. *Pengantar Analisis Data Kategorik*. Yogyakarta: Deepublish.
- [15] Soemartojo, S. M. 2002. *Kajian Metoda CHAID dan CHAID Exhaustive Sebagai Analisa Pohon Bestruktur*. Tesis S2 Program Studi Statistika, Institut Pertanian Bogor.
- [16] Sumartini, S. H. & Purnami, S. W. 2015. *Penggunaan Metode Classification and Regression Tree (CART) untuk Klasifikasi Rekurensi Pasien Kanker Serviks*. Jurnal Sains dan Seni ITS Vol. 4, No. 2, (2015) 2337-3520 (2301-928X Print).
- [17] Susanti, Y., Zukhronah, E., Pratiwi, H., Respatiwan. & Handayani, S. 2017. *Analysis of Chi-square Automatic Interaction Detection (CHAID) and Classification and Regression Tree (CRT) for Classification of Corn Production*. International Conference on Science and Applied Science, 909, doi:10.1088/1742-6596/909/1/012041.
- [18] UCI Machine Learning Repository. *Breast Cancer* (Online), (<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>, diakses 19 Juli 2019).
- [19] Venkatesan, E. & Velmurugan, T. 2015. *Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification*. Indian Journal of Science and Technology, 8(29):1-8,



doi: 10.17485/ijst/2015/v8i29/84646.

- [20] Widiastuti, A. 2001. *Penelusuran Karakteristik Kepemilikan Kartu Kredit Dengan Metode CHAID dan Metode CHAID Exhaustive*. Skripsi S1 Departemen Statistika, Institut Pertanian Bogor.
- [21] Widyastasyah, Y. 2015. *Penentuan Penciri Keragaman Konsumsi Pangan Rumah Tangga Rawan Pangan Dengan Metode CHAID dan Exhaustive CHAID*. Skripsi S1 Departemen Statistika, Institut Pertanian Bogor.
- [22] World Health Organization. 2019. *Cancer* (Online), (<https://www.who.int/health-topics/cancer>, diakses 20 Oktober 2019).
- [23] Yunindya, R. 2017. *Model Credit Scoring Menggunakan Metode Classification and Regression Trees (CART) pada Data Kartu Kredit*. Skripsi S1 Program Studi Statistika, Universitas Islam Bandung.
- [24] Zhang, J., Yu, B. & Chikaraisih, M. 2014. *Interdependences Household Residential and Car Ownership Behavior: A Life History Analysis*. *Journal of Transport Geography* 34 (2014) 165-174.