p-ISSN 2085-7829
e-ISSN 2798-3455

Penerapan Metode Klasifikasi Chi-Square Automatic Interaction Detection dan Exhaustive Chi-Square Automatic Interaction Detection (Studi Kasus: Data Masa Studi Mahasiswa Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Mulawarman)

Application of the Chi-Square Automatic Interaction Detection and Exhaustive Chi-Square Automatic Interaction Detection (Case Study: Study Period Data for Students of the Faculty of Mathematics and Natural Sciences, Mulawarman University)

Nurhasanah[1], Rito Goejantoro[2], dan Suyitno [3]
[1,2]Laboratorium Statistika Komputasi FMIPA Universitas Mulawarman
[3]Laboratorium Statistika Terapan FMIPA Universitas Mulawarman
[1]E-mail: nurhasanah010101@gmail.com

ABSTRACT

The Chi-Square Automatic Interaction Detection (CHAID) and Exhaustive CHAID methods are nonparametric statistical methods that can be used to classify. CHAID and Exhaustive CHAID were used to determine the significant relationship between the dependent variable and the independent variables based on the chi-square independence test. This study was applied to data on the study period of students of FMIPA UNMUL batch 2014. Based on the CHAID and Exhaustive CHIAD methods, it can be seen that the dependent variable of the study period has a significant relationship with the independent variable, namely the study program and GPA predicate. Where students who graduated on time for the Statistics, Biology and Chemistry study program with a satisfactory GPA predicate of 82 students and with a very satisfactory GPA predicate and cum laude with 46 students. Meanwhile, students who did not graduate on time for the Statistics, Biology and Chemistry study program with an adequate GPA predicate of 5 students, a satisfactory GPA predicate of 41 students, very satisfactory and cum laude with 3 students. Students who graduated on time for the Physics study program were 13 students and those who did not graduate on time were 34 students. The chi-square independence test performed on the CHAID method uses fewer possible categorical pairs than the Exhaustive CHAID method which uses all possible categorical pairs so that it requires a long computational and calculation time.

Keywords: *CHAID, exhaustive CHAID, classification*

Pendahuluan

Klasifikasi merupakan proses menemukan model berupa pohon keputusan dengan maksud untuk mendapatkan perkiraan suatu objek yang belum diketahui labelnya. Klasifikasi dibedakan menjadi dua kelompok, yaitu nonparametrik dan parametrik. Adapun kelompok parametrik antara lain regresi logistik dan diskriminan yang memerlukan asumsi. Sedangkan kelompok nonparametrik antara lain *Classification and Regression Tress* (CART), *Chi-Square Automatic Interaction Detection* (CHAID), *Exhaustive* CHAID, dan lain-lain yang tidak memerlukan asumsi (Sulviana, Wigena dan Indahwati, 2018).

CHAID adalah sebuah metode untuk mengklasifikasikan data kategori di mana tujuan dari prosedurnya adalah untuk membagi rangkaian data menjadi subgrup-subgrup berdasarkan pada variabel dependennya.

Bagian utama dari metode CHAID adalah menggunakan uji *chi-square* dan koreksi Bonferroni, *chi-square* berfungsi untuk menganalisis hubungan antara variabel dependen dengan masing-masing variabel independen. Sedangkan koreksi Bonferroni adalah koreksi yang digunakan ketika beberapa uji statistik dilakukan secara bersamaan dan biasanya digunakan dalam perbandingan ganda (Kunto dan Hasana, 2006). Hasil dari pengklasifikasian dalam CHAID akan ditampilkan dalam sebuah diagram pohon klasifikasi.

Exhaustive CHAID pada dasarnya hampir sama dengan CHAID, perbedaannya terjadi pada algoritma penggabungan.

Metode ini banyak digunakan dalam berbagai bidang, salah satu diantaranya pada bidang pendidikan. Tidak dapat dipungkiri, bahwa perguruan tinggi menjadi salah satu persyaratan dasar dalam mencari pekerjaan. Perguruan tinggi dapat diselesaikan dalam jangka waktu yang telah ditentukan, tentunya sesuai dengan kemauan dan kemampuan mahasiswa. Idealnya mahasiswa dapat menempuh pendidikan S1 selama kurang lebih 4 tahun yaitu 8 semester. Namun pada kenyataannya, tidak sedikit mahasiswa yang menyelesaikan masa studinya melebihi waktu tersebut.

Berdasarkan latar belakang tersebut, maka penulis tertarik untuk melakukan penelitian dengan judul "Penerapan Metode Klasifikasi *Chi-Square automatic Interaction Detection* (CHAID) dan *Exhaustive* CHAID pada

p-ISSN 2085-7829
e-ISSN 2798-3455

Mahasiswa Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman Angkatan 2014".

Metode CHAID

CHAID merupakan salah satu teknik iteratif yang menguji satu persatu variabel independen yang digunakan dalam klasifikasi, dan menyusunnya berdasarkan tingkat signifikansi uji *Chi-Square* terhadap variabel dependen (Gallagher, 2000). Terdapat algoritma pada CHAID dimana digunakan untuk melakukan pemisahan dan penggabungan kategori-kategori dalam variabel yang dipakai dalam analisisnya. Ada tiga tahap dalam menganalisis algoritma CHAID yaitu sebagai berikut:

1. Penggabungan (*Merging*)

Tahap pertama dalam algoritma CHAID adalah penggabungan (*merging*). Pada tahap ini akan diperiksa signifikansi dari masing-masing dari kategori variabel independen terhadap variabel dependen. Tahap penggabungan untuk setiap variabel independen dalam menggabungkan kategori-kategori yang tidak signifikan adalah

a) Membuat tabel kontingensi dua arah untuk masing-masing variabel independen dengan variabel dependen
b) Menghitung uji *chi-square* untuk setiap pasangan kategori yang dapat dipilih untuk digabungkan menjadi satu. langkah-langkah dalam uji hipotesis *Chi-Square* adalah sebagai berikut
   i) Menuliskan hipotesis
      H_0: tidak terdapat hubungan antara kedua variabel
      H_1: terdapat hubungan antara kedua variabel
   ii) Menghitung statistik uji
      Perhitungan nilai χ^2_{hitung} untuk tabel kontingensi 2×2 diperoleh dari persamaan koreksi Yates. Persamaan koreksi Yates adalah sebagai berikut

$$\chi^2_{hitung} = \frac{N((n_{11}n_{22}-n_{12}n_{21})-\frac{N}{2})^2}{(n_{11}+n_{12})(n_{11}+n_{21})(n_{12}+n_{22})(n_{21}+n_{22})} \quad (1)$$

      Untuk tabel kontingensi $r \times c$, perhitungan nilai χ^2_{hitung} diperoleh dari persamaan berikut

$$\chi^2_{hitung} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-p_{ij})^2}{p_{ij}} \quad (2)$$

      untuk menghitung frekuensi harapan masing-masing sel digunakan rumus sebagai berikut

$$p_{ij} = \frac{n_{i\bullet}\times n_{\bullet j}}{N} \quad (3)$$

      dengan,
      i=1,2,…,r dan *j=1,2,…,c*
      Dimana $n_{i\bullet}$ adalah total banyaknya pengmatan pada baris ke-*i*, $n_{\bullet j}$ adalah total banyaknya pengamatan pada kolom ke-*j* dan *N* adalah jumlah data dibawah baris *i* dan kolom *j*.
   iii) Kriteria uji
      Keputusan diambil dengan melihat nilai χ^2_{hitung} yang kemudian akan dibandingkan dengan $\chi^2_{(\alpha,(r-1)(c-1))}$, dimana keputusan dapat berupa menolak H_0 atau gagal menolak H_0.
c) Gabungkan pasangan kategori yang tidak signifikan (yaitu pasangan yang mempunyai nilai *chi-square* terbesar dan nilai *p-value* terkecil) menjadi sebuah kategori tunggal.
d) Memeriksa kembali signifikansi kategori baru setelah digabung dengan kategori lainnya dalam variabel independen. Jika masih ada pasangan yang belum signifikan, ulangi langkah c). Jika semua sudah signifikan, lanjutkan langkah berikutnya
e) Menghitung nilai *p-value* terkoreksi Bonferroni berdasarkan pada pasangan kategori yang telah digabung. Gallagher (2000) menyebutkan bahwa pengali Bonferroni untuk masing-masing jenis variabel independen berbeda yaitu sebagai berikut
   i) Pengali Bonferroni untuk variabel independen monotonik

$$M_m = \binom{c-1}{r-1} = \frac{(c-1)!}{(r-1)!((c-1)-(r-1))!} \quad (4)$$

   ii) Pengali Bonferroni untuk variabel independen bebas

$$M_b = \sum_{i=0}^{r-1}(-1)^i\frac{(r-1)^c}{i!(r-1)!} \quad (5)$$

      dengan M_m adalah pengali Bonferroni untuk variabel independen monotonik, M_b adalah pengali Bonferroni untuk variabel independen bebas, *c* adalah banyaknya kategori variabel independen awal, *r* adalah banyaknya kategori variabel independen setelah penggabungan.

2. Pemisahan (*Splitting*)

Tahap pemisahan memilih variabel independen mana yang akan digunakan sebagai *split node* (pemisah node) yang terbaik. Pemilihan dikerjakan dengan membandingkan nilai *chi-square* (dari tahap

p-ISSN 2085-7829
e-ISSN 2798-3455

penggabungan) pada setiap variabrl independen.

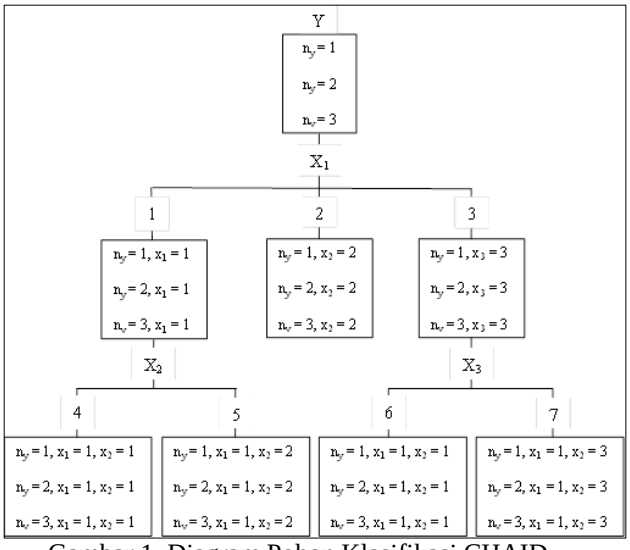
3. Penghentian (*Stopping*)
   Tahap penghentian dilakukan jika proses pertumbuhan pohon harus dihentikan dan tidak ada lagi variabel yang signifikan menunjukkan perbedaan terhadap variabel dependen.

Exhaustive CHAID

Metode *Exhaustive* CHAID dikemukakan oleh Biggs and dkk (1991) yang merupakan evaluasi dari metode sebelumnya yaitu CHAID. Metode *Exhaustive* CHAID mengemukakan penyekatan dengan cara melihat seluruh kemungkinan penggabungan dari pasangan kategori, secara bertahap hingga tersisa dua kategori.

Algoritma *Exhaustive CHAID* sama dengan algoritma *CHAID* yang terdiri dari tiga tahap, yaitu tahap penggabungan (*merging*), pemisahan (*splitting*), dan penghentian (*stopping*). Akan tetapi algoritma *Exhaustive CHAID* melakukan penggabungan lebih menyeluruh sehingga membutuhkan waktu komputasi yang lebih lama daripada *CHAID*.

Menurut Myers (1996), diagram pohon CHAID dan *Exhaustive* CHAID mengikuti aturan dari atas ke bawah (*Top-down stoppig rule*), dimana diagram pohon disusun mulai dari kelompok induk, berlanjut ke bawah sub kelompok yang berturut-turut dari hasil pembagian kelompok induk berdasarkan kriteria tertentu dan tiap node berisi informasi tentang frekuensi variabel *Y*.



Gambar 1. Diagram Pohon Klasifikasi CHAID dan *Exhaustive* CHAID

Metodologi Penelitian

Pada penelitian ini metode yang digunakan untuk penarikan sampel adalah *purposive sampling*. *Purposive sampling* adalah teknik penentuan sampel dengan pertimbangan tertentu (Sugiyono, 2010). Sampling ini dikenal juga sebagai sampling pertimbangan dimana sampel yang digunakan dengan pertimbangan ketersediaan data Universitas Mulawarman.

Variabel Penelitian

Variabel dalam penelitian ini meliputi variabel dependen yaitu masa studi dan variabel independen yaitu jenis kelamin, program studi, jalur masuk universitas dan IPK semester akhir. Berikut adalah devinisi variabel-variabel tersebut

1. Masa Studi
   Masa studi adalah banyaknya semester yang ditempuh oleh mahasiswa. Dalam peraturan akademik UNMUL pada pasal 17, masa studi untuk program sarjana (S1) adalah 8 sampai 14 semester. Jika menempuh lebih dari 8 semester berarti bisa dikatakan melebihi masa studi ideal atau tidak tepat waktu. Akan tetapi pada penelitian ini, peneliti menggunakan masa studi tepat waktu ≤ 10 semester.
2. Jenis Kelamin
   Jenis kelamin mahasiswa dikategorikan laki-laki dan perempuan.
3. Program Studi
   Program Studi dikelompokkan menjadi 4 kategori, yaitu Statistika, Biologi, Kimia dan Fisika.
4. IPK Semester Akhir
   Indeks prestasi kumulatif (IPK) semester akhir akan dijadikan sebagai salah satu variabel independen dengan pembagian menjadi empat kategori berdasarkan peraturan akademik UNMUL pasal 36, berikut pengkategorian IPK :

Tabel 1. Pengkategorian IPK

IPK Semester Akhir	Predikat IPK
2,00 – 2,75	Cukup
2,76 - 3,50	Memuaskan
3,51 - 3,69	Sangat Memuaskan
$\geq$ 3,70	*Cumlaude*

Hasil Penelitian dan Pembahasan

1. Statistika Deskriptif
   Berikut adalah persentase data mahasiswa FMIPA UNMUL angkatan 2014 berdasarkan masa studi, jenis kelamin, program studi, predikat IPK dan jalur masuk universitas dapat dilihat pada Tabel 2. Berdasarkan Tabel 2. untuk nilai *p-value* variabel program studi dikalikan dengan koreksi Bonferroni pada persamaan (5) dan nilai *p-value* variabel predikat IPK dikalikan dengan koreksi Bonferroni pada persamaan (4). Hasil perhitungan koreksi Bonferroni kedua metode tersebut ditunjukkan dalam Tabel 3.

p-ISSN 2085-7829
e-ISSN 2798-3455

1. Tahap Penggabungan Metode CHAID dan *Exhaustive* CHAID

Tahap penggabungan dilakukan menggunakan uji independensi *chi-square* terhadap variabel bebas yaitu jenis kelamin, program studi, predikat IPK dan jalur masuk universitas yang sudah dikategorikan. Tahap ini menghasilkan nilai ststistik *chi-square* dan nilai *p-value* yang ditunjukkan dalam Tabel 2.

Berdasarkan Tabel 3. dilakukan uji independensi *chi-square* pada variabel bebas predikat IPK terhadap kategori program studi statistika; biologi; kimia dan fisika, variabel bebas jenis kelamin untuk program studi statistika; biologi; kimia terhadap kategori predikat IPK cukup, memuaskan dan sangat memuaskan; *cumlaude* dan variabel bebas jenis kelamin untuk program studi fisika terhadap kategori predikat IPK cukup, memuaskan dan sangat memuaskan; *cumlaude* yang bertujuan untuk mengetahui kategori yang menjadi pembagi, dapat dilihat dalam Tabel 4, 5 dan 6.

UNMUL Angkatan 2014

Variabel	Kategori	Persentase
Masa Studi	TW	62,9%
	TTW	37,1%
Jenis Kelamin	Laki-Laki	29,9%
	Perempuan	70,1%
Program Studi	Statistika	17,4%
	Biologi	21%
	Kimia	39,3%
	Fisika	22,3%
Predikat IPK	Cukup	2,2%
	Memuaskan	4,9%
	Sangat Memuaskan	73,7%
	Cumlaude	19,2%
Jalur Masuk Universitas	SNMPTN	17,4%
	SBMPTN	45,1%
	SMMPTN	37,5%

Tabel 2. Persentase Data Mahasiswa FMIPA

Tabel 2. Hasil Pengujian Variabel Terikat dan Variabel Bebas dengan Metode CHAID dan *Exhaustive* CHAID

Variabel Bebas	Kategori	Statistik Uji	Nilai *p-value*	Keputusan
Jenis Kelamin	Laki-laki dan Perempuan	4,067	0,044	H_0 ditolak
Program Studi	Statistika; Biologi; Kimia dan Fisika	31,754	$1{,}75 \times 10^{-8}$	H_0 ditolak
Predikat IPK	Cukup, Memuaskan dan Sangat Memuaskan; *Cumlaude*	30,036	$3{,}005 \times 10^{-7}$	H_0 ditolak

Tabel 3. Hasil Pengujian Variabel Terikat dan Variabel Bebas Terkoreksi Bonferron

Variabel Bebas	Kategori	Statistik Uji	Nilai *p-value*	Keputusan
Jenis Kelamin	Laki-laki dan Perempuan	4,067	0,044	H_0 ditolak
Program Studi	Statistika; Biologi; Kimia dan Fisika	31,754	0^*	H_0 ditolak
Predikat IPK	Cukup, Memuaskan dan Sangat Memuaskan; *Cumlaude*	30,036	$9{,}015 \times 10^{-7*}$	H_0 ditolak

*(nilai *p-value* terkoreksi Bonferroni)

Tabel 4. Nilai Statistik *Chi-Square* Variabel Predikat IPK Terhadap Kategori Program Studi

Variabel Bebas	Program Studi	Statistik Uji	Nilai *p-value*	Keputusan
Predikat IPK	Statistika; Biologi; Kimia	26,401	$1{,}85 \times 10^{-6}$	H_0 ditolak
	Fisika	1,396	0,238	H_0 diterima

p-ISSN 2085-7829
e-ISSN 2798-3455

Tabel 5. Nilai Statistik *Chi-Square* Variabel Jenis Kelamin Terhadap Kategori Predikat IPK

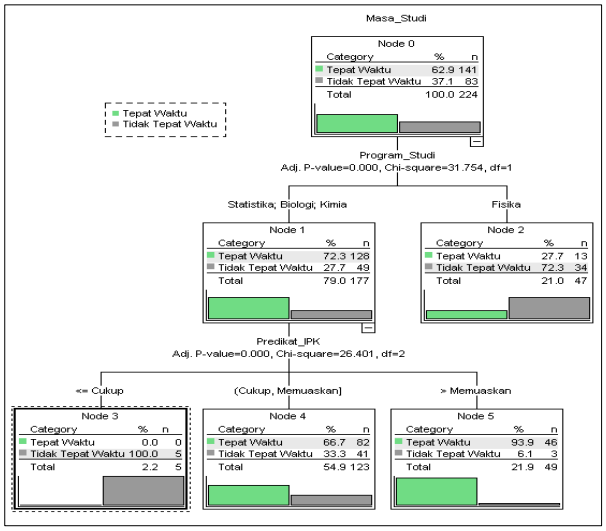
Variabel Bebas	Predikat IPK	Statistik Uji	Nilai *p-value*	Keputusan
Jenis Kelamin	Cukup	0,2	0,655	H_0 diterima
	Memuaskan	2,371	0,124	H_0 diterima
	Sangat Memuaskan; *Cumlaude*	0,265	0,607	H_0 diterima

2. Tahap Pemisahan Metode CHAID dan *Exhaustive* CHAID

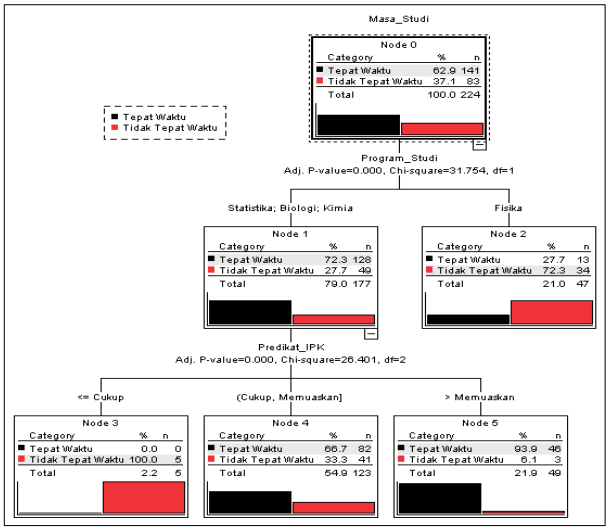
Tahap pemisahan pada metode CHAID dan *Exhaustive* CHAID memilih variabel bebas yang paling berhubungan signifikan terhadap masa studi. Variabel bebas tersebut adalah program studi dan Predikat IPK.

3. Tahap Penghentian Metode CHAID dan *Exhaustive* CHAID

Tahap penghentian dilakukan apabila tidak ada lagi variabel bebas lainnya yang berhubungan signifikan terhadap masa studi. Hasil akhir tahap ini berupa diagram pohon klasifikasi yang ditunjukkan pada Gambar 2. dan Gambar 3. sebagai berikut



Gambar 2. Diagram Pohon Klasifikasi CHAID



Gambar 3. Diagram Pohon Klasifikasi *Exhaustive* CHAID

Tabel 6. Pengklasifikasin Data Mahasiswa FMIPA UNMUL Angkatan 2014

Klasifikasi	**Node**	**Karakteristik**
		Mahasiswa FMIPA UNMUL Angkatan 2014
1	1,3	Mahasiswa program studi Statistika, Biologi dan Kimia dengan predikat IPK cukup
2	1,4	Mahasiswa program studi Statistika, Biologi dan Kimia dengan predikat IPK memuaskan
3	1,5	Mahasiswa program studi Statistika, Biologi dan Kimia dengan predikat IPK sangat memuaskan dan *cumlaude*
4	2	Mahasiswa program studi Fisika

p-ISSN 2085-7829
e-ISSN 2798-3455

Berdasarkan Tabel 6. Diperoleh bahwa hasil klasifikasi data masa studi mahasiswa FMIPA UNMUL angkatan 2014 terdiri dari empat tingkatan klasifikasi dengan setiap tingkatan klasifikasi mempunyai node dankarakteristik yang berbeda.

klasifikasi ke-1 mahasiswa program studi Statistika, Biologi dan Kimia dengan predikat IPK cukup, tidak ada mahasiswa yang menempuh studi tepat waktu sedangkan mahasiswa yang menempuh studi tidak tepat waktu yaitu 5 mahasiswa.

Klasifikasi ke-2 mahasiswa program studi Statistika, Biologi dan Kimia dengan predikat IPK memuaskan yang menempuh studi tepat waktu yaitu 82 mahasiswa sedangkan mahasiswa yang menempuh studi tidak tepat waktu yaitu 41 mahasiswa.

Klasifikasi ke-3 mahasiswa program studi Statistika, Biologi dan Kimia dengan predikat IPK sangat memuaskan; *cumlaude* yang menempuh studi tepat waktu yaitu 46 mahasiswa sedangkan mahasiswa yang menempuh studi tidak tepat waktu yaitu 3 mahasiswa.

Klasifikasi ke-4 mahasiswa program studi Fisika yang menempuh studi tepat waktu yaitu 13 mahasiswa dan mahasiswa yang menempuh studi tidak tepat waktu yaitu 34 mahasiswa .

Klasifikasi ke-2 mahasiswa program studi Statistika, Biologi dan Kimia dengan predikat IPK memuaskan yang menempuh studi tepat waktu yaitu 82 mahasiswa sedangkan mahasiswa yang menempuh studi tidak tepat waktu yaitu 41 mahasiswa.

Klasifikasi ke-3 mahasiswa program studi Statistika, Biologi dan Kimia dengan predikat IPK sangat memuaskan; *cumlaude* yang menempuh studi tepat waktu yaitu 46 mahasiswa sedangkan mahasiswa yang menempuh studi tidak tepat waktu yaitu 3 mahasiswa.

Klasifikasi ke-4 mahasiswa program studi Fisika yang menempuh studi tepat waktu yaitu 13 mahasiswa dan mahasiswa yang menempuh studi tidak tepat waktu yaitu 34 mahasiswa .

Kesimpulan

Berdasarkan uraian di atas, maka kesimpulan dari penelitian ini adalah:

1. Berdasarkan metode *Chi-Square Automatic Interaction Detection* (CHAID), variabel yang berpengaruh signifikan pada masa studi mahasiswa FMIPA UNMUL angkatan 2014 adalah variabel adalah variabel program studi dan predikat IPK.
2. Berdasarkan metode *Exhaustive CHAID* variabel yang berpengaruh signifikan pada masa studi mahasiswa FMIPA UNMUL angkatan 2014 adalah variabel program studi dan predikat IPK .
3. Metode yang lebih baik pada penelitian ini adalah metode CHAID karena pada uji independensi menggunakan kemungkinan pasangan kategori yang lebih sedikit dibandingkan dengan metode *Exhaustive* CHAID yang menggunakan seluruh kemungkinan pasangan kategori sehingga membutuhkan waktu komputasi dan perhitungan yang lama.

Daftar Pustaka

Biggs, D., Ville, B., & Suen, E. (1991). A Method of Choosing Multiway Partitions for Classification and Decision Trees. *Journal of Applied Statistics*, 18 (1), 49-62.

Gallagher, C.A. (2000). *An Iterative Approach to Classification Analysis*. http://casualyactuaries.com/pubs/dpp90/90dpp237.pdf. Tanggal Akses : 13 Desember 2019.

Kunto, Y.S., dan Hasana, S.N. (2006). "Analisis CHAID Sebagai Alat Bantu Statistika Untuk Segmetasi Pasar (Studi Kasus pada Koperasi Syari'ah Al-Hidayah)". *Jurnal Manajemen Pemasaran*, 2 (2), 88-98.

Myers, J.H. (1996). *Segmentation and positioning for Stategic Marketing Decicions*. Chichago: American Marketing Association.

Sugiyono. (2010). *Statistika Untuk Penelitian*. Bandung: Alfabeta.

Sulviana, V., Wigena, A.H., dan Indahwati. (2018). "Implementasi Metode CHAID (*Chi-Square Automatic Interaction Detection*) pada Segmentasi *Trand* Penjualan Minuman Ringan di Indonesi". *Xplore*, 2 (2), 24-31.