ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 14, Nomor 2, Tahun 2025, Halaman 489 - 499
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PENGKLASIFIKASIAN PENYAKIT HIPERTENSI MENGGUNAKAN METODE *CHI-SQUARE AUTOMATIC INTERACTION DETECTION* (Studi Kasus: Pengunjung Kegiatan Kaltim Expo 2023)

Kharisma Dwi Wahyuni[1], Darnah[2*], M. Fathurahman[3]
[1,2,3] Program Studi Statistika, Universitas Mulawarman
***e-mail**: darnah.98@gmail.com

DOI: 10.14710/j.gauss.14.2.489-499

Article Info:
Received: 2025-06-05
Accepted: 2025-11-12
Available Online: 2025-11-19

Keywords:
CHAID; hypertension; classification; decision tree

Abstract: Data mining is a technology used as an automated tool in the decision-making with classification being one of its techniques. The Chi-square Automatic Interaction Detection method classifies data by dividing samples into groups based on certain criteria, displaying results in a tree diagram. This study aims to obtain related factors, classification results, and accuracy of hypertension status classification of visitors to the East Kalimantan Provincial Health Office stand at the 2023 Kaltim Expo in Samarinda City using the CHAD method. The study found that age and family history are factors related to hypertension status. Classification results showed 10 elderly visitors without a family history of hypertension (2 with hypertension, 8 without), 216 non-elderly visitors without a family history of hypertension (9 with hypertension, 207 without), 10 elderly visitors with a family history of hypertension (9 with hypertension, 1 without), and 117 non-elderly visitors with a family history of hypertension (36 with hypertension, 81 without). The accuracy of hypertension status classification using the CHAID method was 86%.

1. PENDAHULUAN

Data mining adalah teknologi yang sangat berguna sebagai alat bantu otomatis dalam proses penetapan kebijakan dan pengambilan keputusan. Tujuan utama dari *data mining* yaitu digunakan untuk mengekstrak pola, informasi yang bermanfaat dari data yang besar. Teknik yang digunakan dalam *data mining* salah satunya adalah klasifikasi yang merupakan proses penciptaan model atau fungsi yang mengidentifikasi dan membedakan kelas data sehingga dipakai untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Han & Kamber, 2006).

Metode dalam teknik klasifikasi salah satunya adalah metode *Chi-square Automatic Interaction Detection* (CHAID) yang diperkenalkan oleh Dr. G. V. Kass pada tahun 1980 dalam sebuah artikel "*An Exploratory Technique for Investigating Large Quantities of Categorial Data*" yang dimuat dalam buku *Applied Statistics*. Menurut Everit dan Skrondal (2010), metode CHAID ini sering disebut sebagai metode pohon klasifikasi. Metode ini bertujuan untuk membagi data menjadi kelompok-kelompok yang kecil berdasarkan hubungan antara variabel dependen dan independen. Analisis ini digunakan pada data dengan variabel kategorik yang memberikan label sesuai dengan pengamatan dan dikelompokkan ke dalam beberapa kategori. Hasil dari pengklasifikasian CHAID akan disajikan dalam bentuk diagram pohon (Lehmann & Eherler, 2001).

Penerapan metode CHAID melibatkan pohon keputusan dari akar pohon ke daun yang memberikan hasil klasifikasi, yaitu *decision tree* yang merinci proses pengambilan keputusan dalam suatu struktur pohon. Menurut Cho dan Ngai (2003), pohon keputusan atau *decision tree* adalah cara yang dipakai untuk mengklasifikasikan objek dalam bentuk gambar, daun, dan cabang. Daun (*leaf*) pada *decision tree* menunjukan kelas dari data yang

ada, dan cabangnya (*internal node*) menunjukan kondisi dari atribut objek yang terukur. Pohon keputusan atau *decision tree* memiliki akar di bagian atas dan daunnya berada di bawah akan terlihat seperti pohon terbalik. Metode CHAID sangat efektif dalam menganalisis data dengan jumlah besar yang seluruh variabelnya bersifat kategorik. Hasil analisis menjadi lebih mudah dipahami dan dapat mengidentifikasi faktor paling signifikan diantara faktor lainnya ketika menggunakan metode CHAID. Metode CHAID adalah salah satu metode untuk mengklasifikasikan data secara sistematis dan juga terstruktur (Miftahuddin, 2012).

Penerapan metode CHAID dapat digunakan pada bidang kesehatan, seperti mengelompokkan pasien yang terkena penyakit hipertensi dan tidak terkena hipertensi. Menurut WHO dalam Sujana dan Trisyan (2023), hipertensi adalah masalah kesehatan yang serius di seluruh dunia karena dapat meningkatkan risiko terkena penyakit kardiovaskuler. Pada tahun 2016, penyakit jantung iskemik dan stroke menjadi penyebab kematian utama di dunia. Berdasarkan data dari Dinas Kesehatan Kota Samarinda, penderita hipertensi di Samarinda pada tahun 2021 yaitu berjumlah 33.085 kasus pada semua kalangan umur dengan prevalensi 24,9% (Dinas Kesehatan Kota Samarinda, 2021). Sedangkan berdasarkan data dari Dinas Kesehatan Provinsi Kalimantan Timur, penderita hipertensi di Samarinda pada tahun 2022 berjumlah 11.995 kasus (Dinas Kesehatan Provinsi Kalimantan Timur, 2022). Hal ini menunjukkan pentingnya penelitian yang dapat membantu dalam mengklasifikasikan dan memprediksi risiko hipertensi secara akurat, karena menurut WHO hipertensi adalah masalah kesehatan yang serius di seluruh dunia. Data hipertensi perlu diteliti agar dapat digunakan secara optimal dalam upaya pengendalian penyakit ini.

Terdapat beberapa penelitian sebelumnya yang menggunakan metode CHAID. Rizki, Umam, dan Hamzah (2020) melakukan penelitian untuk mengklasifikasikan nasabah berdasarkan status kredit nasabah sebagai variabel dependen dan data pribadi nasabah sebagai variabel independen. Dari 7 variabel independen yang diuji menggunakan uji *Chi-square*, hanya 5 variabel yang signifikan terhadap variabel dependen. Hasil analisis CHAID menunjukkan adanya empat kelas. Selain itu Fanggidae, dkk (2021) juga melakukan penelitian untuk mengklasifikasikan faktor-faktor yang mempengaruhi prestasi akademik mahasiswa Pendidikan Matematika FKIP Universitas Nusa Cendana dengan metode CHAID, menghasilkan diagram pohon terdapat dua variabel yang berhubungan yaitu rata-rata nilai UN mahasiswa dan jalur masuk mahasiswa.

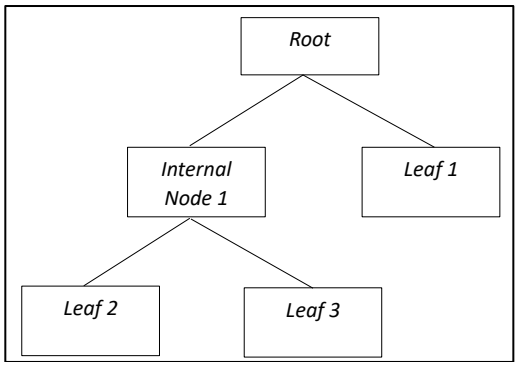
2. TINJAUAN PUSTAKA

Data mining adalah sebuah proses yang bertujuan untuk menggali pengetahuan atau pola dari kumpulan data yang besar (Witten dkk, 2011). Suyanto (2017) mendefinisikannya sebagai kombinasi berbagai teknik ilmu komputer yang digunakan untuk menemukan pola baru dalam kumpulan data besar. Proses ini menggabungkan berbagai metode dari kecerdasan buatan, *machine learning*, statistik, dan sistem basis data. Secara umum, proses *data mining* terbagi menjadi enam tahap. Tiga tahap awal yaitu pembersihan data (*data cleaning*), dan transfomasi data (*data transformation*), dikenal sebagai proses praproses data. Tahap selanjutnya meliputi penggalian data (*data mining*), evaluasi pola (*pattern evaluation*), dan penyajian pola (*knowledge presentation*) (Silitonga, 2016).

Klasifikasi adalah proses pencarian pola dalam data untuk membuat model berupa pohon keputusan dengan tujuan memprediksi label dari objek yang belum diketahui. Klasifikasi juga diartikan sebagai proses pengelompokan suatu hal menjadi beberapa kelompok berdasarkan persamaan dan perbedaannya. Masalah klasifikasi sering ditemui pada kehidupan sehari-hari, baik di bidang pendidikan, sosial, industri, kesehatan maupun perbankan (Agresti, 1990). Tujuan utama dari klasifikasi adalah untuk melakukan prediksi

terhadap kategori data yang diinput. Dalam *data mining* ada berbagai teknik klasifikasi yang digunakan dalam mengelompokkan data ke dalam kategori tertentu. Teknik-teknik ini antara lain seperti teknik pohon keputusan, bayesian, jaringan saraf tiruan, aturan asosiasi, dan teknik lainnya (Hamidah, 2013).

Menurut Cho dan Ngai (2003), pohon keputusan digunakan untuk mengelompokkan objek berdasarkan kondisi yang biasanya direpresentasikan melalui gambar daun dan cabang. Daun (*leaf*) mewakili kelas objek, sedangkan cabangnya *(internal node*) menunjukkan kondisi atribut objek yang dapat diukur. Pada sebuah pohon keputusan, level teratas adalah akar (*root*) yang paling menentukan kelas objek. Konsep dasar pohon keputusan dapat dilihat pada Gambar 1.



Gambar 1. Konsep Dasar Pohon Keputusan (Nugi, 2012)

Walpole (1995) menjelaskan bahwa uji *Chi-square* digunakan untuk mengetahui apakah data pengamatan (*observed*) sesuai dengan data harapan (*expected*). Uji *Chi-square* juga dapat digunakan untuk menentukan apakah karakteristik populasi adalah sama atau seragam (homogen). Beberapa syarat uji independensi *Chi-square* menurut Aminoto dan Agustina (2020), yaitu sebagai berikut:

1. Tidak ada sel dengan nilai frekuensi harapan sebesar nol.
2. Pada tabel kontingensi 2×2, tidak boleh ada satu sel dengan frekuensi harapan kurang dari 5.
3. Tidak boleh lebih dari 20% dari total sel untuk jumlah sel dengan frekuensi harapan kurang dari 5 pada tabel kontingensi dengan ukuran lebih dari 2×2, misal 2×3.

Jika tabel kontingensi 2×2 memenuhi persyaratan diatas, maka yang digunakan adalah koreksi Yates. Apabila tabel kontingensi 2×2 tetapi tidak memenuhi syarat diatas, maka menggunakan rumus *Fisher Exact Test*. Untuk tabel kontingensi lebih dari 2×2 menggunakan *Pearson Chi-square*.

Ketentuan penggunaan uji *Chi-square* untuk tabel kontingensi 2×2 berdasarkan ukuran sampel menurut Sugiyono (2015) adalah sebagai berikut (Aminoto & Agustina, 2020):

1. Jika jumlah sampel > 40, menggunakan uji *Chi-square* dengan koreksi Yates dalam kondisi apapun.
2. Jika jumlah sampel antara $20 - 40$ dengan catatan:
   - Jika tidak terdapat nilai frekuensi harapan < 5, menggunakan uji *Chi-square* dengan koreksi Yates.
   - Jika terdapat nilai frekuensi harapan < 5, menggunakan *Fisher Exact Test.*
3. Jika jumlah sampel < 20, menggunakan *Fisher Exact Test.*

   Statistik uji *Pearson Chi-square* adalah:

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

Frekuensi harapan masing-masing sel dapat dihitung dengan mengunakan rumus sebagai berikut:

$$E_{ij} = \frac{n_i.\times n._j}{n} \quad (2)$$

Statistik uji *Fisher Exact Test* adalah sebagai berikut:

$$\chi^2 = \frac{(n_{11}+n_{21})!\,(n_{12}+n_{22})!\,(n_{11}+n_{12})!\,(n_{21}+n_{22})!}{n!\,n_{11}!\,n_{21}!\,n_{12}!\,n_{22}!} \quad (3)$$

Statistik uji *Chi-square* dengan koreksi Yates sebagai berikut:

$$\chi^2 = \frac{n\left(|n_{11}n_{22} - n_{12}n_{21}| - \frac{n}{2}\right)^2}{(n_{11}+n_{12})(n_{11}+n_{21})(n_{12}+n_{22})(n_{21}+n_{22})} \quad (4)$$

dengan:

χ^2 : Statistik hitung *Chi-square*

n_{11} : Banyak pengamatan dengan sifat kategori 1 pada variabel A (A_1) dan kategori 1 pada variabel B (B_1)

n_{12} : Banyak pengamatan dengan sifat kategori 1 pada variabel A (A_1) dan kategori 2 pada variabel B (B_2)

n_{21} : Banyak pengamatan dengan sifat kategori 2 pada variabel A (A_2) dan kategori 1 pada variabel B (B_1)

n_{22} : Banyak pengamatan dengan sifat kategori 2 pada variabel A (A_2) dan kategori 2 pada variabel B (B_2)

n : Banyaknya seluruh pengamatan.

Hipotesis *Chi-square* yang digunakan adalah sebagai berikut:

H_0 : $p = 0$
(Tidak terdapat hubungan antara variabel pertama dan variabel kedua)

H_1 : $p \neq 0$
(Terdapat hubungan antara variabel pertama dan variabel kedua)

Dasar pengambilan keputusan hipotesis berdasarkan perbandingan χ^2_{hitung} dengan $\chi^2_{tabel(\alpha;(b-1)(k-1))}$ yaitu jika $\chi^2_{hitung} \geq \chi^2_{tabel(\alpha;(b-1)(k-1))}$, maka H_0 ditolak. Sedangkan dasar pengambilan keputusan hipotesis berdasarkan taraf signifikansi yaitu jika p-*value* $\leq \alpha$, maka H_0 ditolak (Santoso, 2012).

Metode CHAID digunakan untuk mengklasifikasikan data kategori dengan tujuan membagi data ke dalam kelompok berdasarkan variabel dependen. Hasil dari klasifikasi menggunakan metode CHAID akan ditampilkan dalam diagram pohon (Lehmann & Eherler, 2001). Menurut Gallagher, dkk (2000), CHAID memiliki tiga bentuk yang berbeda untuk membagi variabel bebas kategorik, yaitu:

1. Monotonik (*Monotonic*)

Penggabungan kategori terjadi ketika variabel berdekatan, seperti usia dan pendapatan yang mengikuti urutan asli (data ordinal).

2. Bebas (*Free*)

Variabel ini memungkinkan kategori-kategori untuk dikombinasi atau digabung, baik ketika berdekatan maupun tidak (data nominal), seperti pekerjaan, etnis, atau wilayah geografis.

3. Mengambang (*Floating*)

Variabel ini dianggap monotonik kecuali untuk kategori yang *missing value* yang dapat dikombinasikan dengan kategori manapun.

Koreksi Bonferroni merupakan proses mengoreksi yang dipakai untuk mengontrol tingkat signifikansi atau resiko kesalahan ketika uji statistik dilakukan pada waktu yang bersamaan (Kunto & Hasana, 2006). Gallagher, dkk (2000) mengemukakan bahwa pengali Bonferroni untuk setiap jenis variabel independen adalah sebagai berikut:

1. Monotonik

$$M = \binom{(c-1)}{(r-1)} \tag{5}$$

2. Bebas

$$M = \sum_{m=0}^{r-1} (-1)^m \frac{(r-1)^c}{m!\,(r-1)!} \tag{6}$$

3. Mengambang

$$M = \binom{c-2}{r-2} + r\binom{c-2}{r-2} \tag{7}$$

dengan:

M = Pengali Bonferroni

c = Banyaknya kategori variabel independen awal

r = Banyaknya kategori variabel independen setelah penggabungan

Diagram pohon dimulai dari *root node* (*node* akar) dan melalui tiga tahap pada setiap simpul berulang kali sebagai berikut (Wulandary, 2014):

1. Tahap Penggabungan (*Merging)*

Proses penggabungan digunakan untuk menggabungkan kategori–kategori pada suatu variabel independen yang memiliki lebih dari dua kategori. Tahap penggabungan untuk setiap variabel independen yang memiliki lebih dari dua kategori dalam menggabungkan kategori adalah sebagai berikut:

   a. Membentuk tabel kontingensi dua arah dari sebuah pasangan kategori yang dapat digabung menjadi kategori tunggal.
   b. Menghitung statistik *Chi-square* untuk setiap pasangan kategori yang memungkinkan untuk digabungkan menjadi satu dengan hipotesis sebagai berikut:
      $H_0 : p = 0$
      (Tidak terdapat hubungan antara variabel pertama dan variabel kedua)
      $H_1 : p \neq 0$
      (Terdapat hubungan antara variabel pertama dan variabel kedua)
      Keputusan yang diambil dari uji *Chi-square* ini adalah H_0 ditolak jika nilai $\chi^2_{hitung} \geq \chi^2_{tabel(\alpha;(b-1)(k-1)}$ atau p-*value* $\leq \alpha$.
   c. Memeriksa kembali signifikansi kategori baru setelah digabungkan dengan kategori lain dalam variabel independen. Jika kategori tidak signifikan, maka variabel tersebut tidak dapat digunakan ke tahap selanjutnya. Jika kategori tersebut signifikan setelah penggabungan, maka dapat digunakan ke tahap selanjutnya yaitu koreksi Bonferroni.
   d. Menghitung p-*value* terkoreksi Bonferroni dari variabel independen yang signifikan setelah penggabungan.

2. Tahap Pemisahan (*Splitting)*

Tahap *splitting* melibatkan pemilihan variabel independen yang akan digunakan sebagai *split node* (pemisah *node*) terbaik. Variabel dipilih berdasarkan perbandingan p-*value* atau

statistik hitung dari tahap *merging*. Variabel yang memiliki p-*value* terkecil dan statistik hitung terbesar kedua setelah variabel independen digunakan sebagai akar.

3. Tahap Penghentian (*Stopping)*

Tahap *stopping* dilakukan ketika pertumbuhan pohon harus dihentikan sesuai dengan peraturan penghentian sebagai berikut:

a. Tidak ada lagi variabel independen yang signifikan.
b. Jika pohon mencapai batas nilai pohon maksimum dari spesifikasi, proses pertumbuhan akan berhenti.

Hasil klasifikasi dapat dievaluasi kinerja klasifikasinya dengan mengukur *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Matriks yang menunjukkan TP, TN, FP, dan FN, ini sering disebut sebagai *confusion matrix* yang dapat dilihat pada Tabel 1.

Tabel 1. *Confusion Matrix*

	***Positive* (Prediksi)**	***Negative* (Prediksi)**
***Positive* (Aktual)**	TP	FN
***Negative* (Aktual)**	FP	TN

Berdasarkan nilai TP, TN, FP, dan FN, dapat diperoleh nilai akurasi yang digunakan untuk mengetahui tingkat keakuratan dari hasil klasifikasi sebagai berikut (Tyas, 2015):

$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad (8)$$

3. METODE PENELITIAN

Populasi yang digunakan dalam penelitian ini yaitu pengunjung *stand* Dinas Kesehatan Provinsi Kalimantan Timur pada kegiatan Kaltim Expo tahun 2023. Sampel yang digunakan dalam penelitian ini merupakan pengunjung *stand* Dinas Kesehatan Provinsi Kalimantan Timur pada kegiatan Kaltim Expo tahun 2023 yang berdomisili di Kota Samarinda dengan jumlah sampel sebanyak 353 orang. Penelitian ini menggunakan ketersediaan data dari formulir Dinas Kesehatan Provinsi Kalimantan Timur dengan 6 variabel yang dapat dilihat pada Tabel 2.

Tabel 2. Variabel Penelitian

Notasi	Nama Variabel	Definisi Operasional	Skala Pengukuran Variabel	Kategori
Y	Hipertensi	Status penyakit hipertensi pada pengunjung yang diamati dari formulir pemeriksaan kesehatan	Nominal	0 = Tidak 1 = Ya
X_1	Usia	Usia pengunjung pada saat pemeriksaan kesehatan	Nominal	0 = Lansia (> 60 tahun) 1 = Bukan lansia (20 ≤ 60 tahun)
X_2	Jenis kelamin	Jenis kelamin pengunjung	Nominal	0 = Laki-laki 1 = Perempuan
X_3	Riwayat keluarga	Kondisi apakah keluarga pengunjung terdapat riwayat hipertensi atau tidak	Nominal	0 = Tidak 1 = Ya
X_4	Konsumsi rokok	Kondisi pengunjung berdasarkan kebiasaan merokok yang diamati dari formulir pemeriksaan kesehatan	Nominal	0 = Tidak 1 = Ya

X_5	Indeks Massa Tubuh	Status gizi pengunjung yang dihitung menggunakan data berat badan dan tinggi badan dari formulir pemeriksaan kesehatan	Ordinal	1 = Kurus 2 = Normal 3 = Gemuk

Terkait dengan isu-isu dan tujuan penelitian ini, teknik analisis data yang digunakan adalah sebagai berikut:

1. Tahap penggabungan (*merging*) untuk variabel yang memiliki lebih dari dua kategori dimana dalam penelitian ini yaitu variabel IMT (X_5).
2. Tahap pemisahan (*splitting*) untuk memilih variabel independen yang akan digunakan untuk membagi *node.*
3. Tahap penghentian (*stopping*) setelah tidak ada lagi variabel independen yang signifikan.
4. Penarikan kesimpulan.

4. HASIL DAN PEMBAHASAN

Langkah analisis data yang pertama adalah tahap penggabungan untuk variabel yang memiliki lebih dari dua kategori dimana dalam penelitian ini yaitu variabel IMT (X_5). Jika dalam suatu variabel independen hanya terdapat dua kategori maka tidak perlu ada penggabungan. IMT termasuk ke dalam skala ordinal yang memperhatikan urutan kategori sehingga tidak bisa menggabungkan kategori secara acak. Berdasarkan uji independensi *Chi-square* yang telah dilakukan dari tahap penggabungan, dapat dirangkum hasil pengujian sebagai berikut:

Tabel 3. Hasil Uji Independensi *Chi-square* Status Penyakit Hipertensi dan IMT

Variabel Dependen	IMT	χ^2_{hitung}	*p-value*	Keputusan
Status Penyakit Hipertensi	Kurus dan Normal	0	1	H_0 gagal ditolak
	Normal dan Gemuk	1,31	0,2519	H_0 gagal ditolak

Pada Tabel 3, dapat dilihat bahwa terdapat dua pasangan kategori yang tidak berhubungan signifikan. Dari kedua pasangan kategori tersebut yang memiliki nilai statistik uji terbesar dan nilai *p-value* terkecil adalah IMT normal dan gemuk sehingga kategori IMT normal dan gemuk digabung menjadi satu kategori tunggal namun tetap tidak berhubungan signifikan sehingga variabel IMT tidak dilanjutkan ke tahap berikutnya yaitu koreksi Bonferroni dan variabel IMT tidak termasuk ke dalam pohon keputusan.

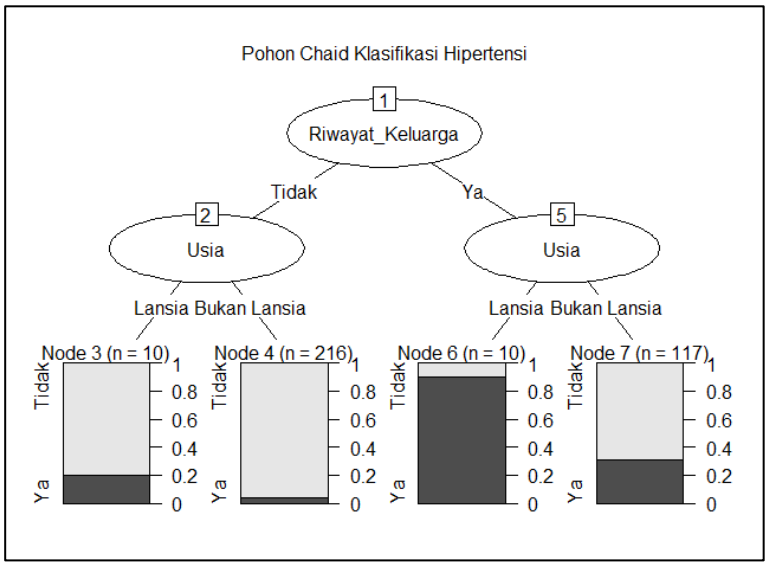
Selanjutnya tahap pemisahan yaitu memilih variabel independen yang akan digunakan untuk membagi *node.* Pemilihan dilakukan dengan memilih variabel independen yang memiliki nilai *Chi-square* terbesar kedua setelah variabel independen yang digunakan sebagai simpul induk pada pohon keputusan. Hasil uji independensi *Chi-square* variabel dependen dan variabel independen yang signifikan setelah tahap penggabungan dapat dirangkum pada Tabel 4.

Tabel 4. Hasil Pengujian Uji Independensi *Chi-square* Variabel Dependen dan Variabel Independen (setelah tahap penggabungan)

Variabel Independen	Kategori	χ^2_{hitung}	*p-value*	Keputusan
Usia	0 = Lansia 1 = Bukan Lansia	21,32	$3{,}887 \times 10^{-6}$	H_0 ditolak
Riwayat Keluarga	0 = Tidak 1 = Ya	54,64	$1{,}443 \times 10^{-13}$	H_0 ditolak

Pada Tabel 4, dapat dilihat bahwa urutan nilai statistik uji *Chi-square* terbesar dan nilai *p-value* terkecil pertama adalah variabel riwayat keluarga dan yang kedua adalah variabel usia. Berdasarkan urutan tingkat signifikansi variabel independen tersebut, maka variabel riwayat keluarga digunakan sebagai simpul induk pada pohon keputusan (*node* 1) dengan kategori tidak dan kategori ya.

Tahap penghentian dilakukan setelah tidak ada lagi variabel independen yang signifikan. Hasil akhir dari langkah ini adalah diagram pohon klasifikasi yang dapat dilihat pada Gambar 2.



Gambar 2. Diagram Pohon Klasifikasi Status Penyakit Hipertensi

Pada Gambar 2 tampak bahwa ada dua dari lima variabel independen yang berhubungan signifikan dengan variabel dependen, yaitu riwayat keluarga dan usia, sehingga ada tiga variabel independen yang dinyatakan tidak mempunyai hubungan signifikan dengan variabel dependen, yaitu jenis kelamin, konsumsi rokok, dan IMT. Pada hasil diagram pohon yang terbentuk dengan metode CHAID terlihat bahwa variabel riwayat keluarga sebagai simpul induk atau *node* 1 karena memiliki nilai statistik *Chi-square* terbesar yaitu 56,91. Selanjutnya, variabel usia termasuk dalam pohon klasifikasi sebagai kedalaman kedua karena memiliki nilai statistik *Chi-square* terbesar kedua yaitu 24,32. Berdasarkan metode CHAID, variabel independen yang tidak berhubungan signifikan dengan variabel dependen setelah diuji menggunakan uji *Chi-square* maka tidak termasuk dalam diagram pohon yang terbentuk, dimana variabel independen yang tidak berhubungan signifikan yaitu jenis kelamin, konsumsi rokok, dan IMT. Pada *node* 3,4,6, dan 7 terjadi proses penghentian karena tidak terdapat variabel signifikan yang dapat menjadi cabang dari *node* tersebut.

Berdasarkan diagram pohon yang terbentuk pada Gambar 2, maka segmentasi pengunjung dapat diklasifikasikan seperti pada Tabel 5.

Tabel 5. Segmentasi Pengunjung

Segmen	Aturan
Segmen 1 (*node* 3)	Jika **riwayat keluarga** tidak memiliki penyakit hipertensi dan **usia** kategori lansia, maka diklasifikasikan **tidak hipertensi**
Segmen 2 (*node* 4)	Jika **riwayat keluarga** tidak memiliki penyakit hipertensi dan **usia** kategori bukan lansia, maka diklasifikasikan **tidak hipertensi**
Segmen 3 (*node* 6)	Jika **riwayat keluarga** memiliki penyakit hipertensi dan **usia** kategori lansia, maka diklasifikasikan **hipertensi**
Segmen 4 (*node* 7)	Jika **riwayat keluarga** memiliki penyakit hipertensi dan **usia** kategori bukan lansia, maka diklasifikasikan **tidak hipertensi**

Hasil prediksi status penyakit hipertensi pada pengunjung *stand* Dinas Kesehatan Provinsi Kalimantan Timur pada kegiatan Kaltim Expo tahun 2023 dapat dilihat pada Tabel 6.

Tabel 6. Hasil Prediksi

Responden	X_1	X_2	X_3	X_4	X_5	Y	Prediksi
1	1	0	1	0	3	Tidak	Tidak
2	1	1	1	0	3	Tidak	Tidak

3	1	1	0	0	3	Ya	Tidak
4	1	1	1	0	3	Tidak	Tidak
5	1	0	0	0	3	Tidak	Tidak
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
353	1	1	1	0	3	Ya	Tidak

Prediksi status penyakit hipertensi pada Tabel 6, hanya memperhatikan variabel yang berhubungan signifikan, yaitu riwayat keluarga (X_3) dan usia (X_1) berdasarkan proporsi yang terbesar pada diagram pohon. Keakuratan klasifikasi dapat diukur menggunakan *confusion matrix* pada Tabel 7.

Tabel 7. *Confusion Matrix*

Hasil Observasi	**Hasil Prediksi**	
	Tidak	**Ya**
Tidak	296	1
Ya	47	9

Akurasi pohon keputusan dalam mengklasifikasikan status penyakit hipertensi dapat dihitung dengan persamaan (8).

$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%$$
$$= \frac{296+9}{296+47+9+1} \times 100\% = 86\%.$$

Berdasarkan nilai akurasi tersebut, maka dapat diketahui bahwa keakuratan metode CHAID dalam mengklasifikasikan status penyakit hipertensi pada pengunjung *stand* Dinas Kesehatan Provinsi Kalimantan Timur pada kegiatan Kaltim Expo tahun 2023 adalah sebesar 86% dan termasuk dalam tingkat akurasi yang sangat baik. Nilai akurasi yang tinggi menunjukkan tingkat kepercayaan terhadap hasil klasifikasi, sehingga hasil penelitian ini dapat digunakan untuk mendukung kebijakan atau tindakan medis dalam penanganan penyakit hipertensi seperti mendeteksi dini risiko hipertensi pada masyarakat. Penelitian selanjutnya diharapkan dapat menggunakan jumlah data yang lebih besar dan membandingkan kinerja CHAID dengan metode klasifikasi lainnya, guna memperoleh model yang optimal dan dapat diimplementasikan di bidang kesehatan.

5. KESIMPULAN

Faktor-faktor yang berhubungan dengan status penyakit hipertensi pada pengunjung *stand* Dinas Kesehatan Provinsi Kalimantan Timur pada kegiatan Kaltim Expo tahun 2023 yang berdomisili di Kota Samarinda berdasarkan metode CHAID adalah riwayat keluarga dan usia.

Hasil klasifikasi status penyakit hipertensi pada pengunjung *stand* Dinas Kesehatan Provinsi Kalimantan Timur pada kegiatan Kaltim Expo tahun 2023 yang berdomisili di Kota Samarinda berdasarkan metode CHAID adalah banyaknya pengunjung *stand* Dinas Kesehatan Provinsi Kalimantan Timur pada kegiatan Kaltim Expo yang tidak mempunyai riwayat keluarga hipertensi dan berusia lansia sebanyak 10 orang yang terdiri dari 2 orang menderita hipertensi dan 8 orang tidak menderita hipertensi. Pengunjung *stand* yang tidak mempunyai riwayat keluarga hipertensi dan berusia bukan lansia sebanyak 216 orang yang terdiri dari 9 orang menderita hipertensi dan 207 orang tidak menderita hipertensi. Sedangkan pengunjung *stand* yang mempunyai riwayat keluarga hipertensi dan berusia lansia sebanyak 10 orang yang terdiri dari 9 orang menderita hipertensi dan 1 orang tidak menderita hipertensi. Pengunjung *stand* yang mempunyai riwayat keluarga hipertensi dan berusia bukan lansia sebanyak 117 orang yang terdiri dari 36 orang menderita hipertensi dan 81 orang tidak menderita hipertensi.

Hasil keakuratan klasifikasi status penyakit hipertensi pengunjung *stand* Dinas Kesehatan Provinsi Kalimantan Timur pada kegiatan Kaltim Expo tahun 2023 yang berdomisili di Kota Samarinda menggunakan metode CHAID adalah sebesar 86%.

DAFTAR PUSTAKA

Aminoto, T. & Agustina, D. (2020). *Mahir Statistika dan SPSS*. Tasikmalaya: Edu Publisher.

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

Cho, V. & Ngai, E. (2003). Data Mining for Selection of Insurance Sales Agents. *Expert Systems: The Journal of Knowledge Engineering*, 20, 123-132.

Dinas Kesehatan Kota Samarinda. (2021). *Profil kesehatan Kota Samarinda tahun 2021*. Samarinda: Dinas Kesehatan Kota Samarinda.

Dinas Kesehatan Provinsi Kalimantan Timur. (2022). *Rencana Strategi (RENSTRA) 2022*. Samarinda: Dinas Kesehatan Provinsi Kalimantan Timur.

Everit, B. S. & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. London: Cambridge University Press.

Fanggidae, J. J. R., Ekowati, C. K., Nenohai, J. M. H., & Udil, P. A. (2021). Klasifikasi faktor-faktor yang mempengaruhi prestasi akademik mahasiswa pendidikan matematika FKIP UNDANA dengan metode CHAID. *Fraktal: Jurnal Matematika Dan Pendidikan Matematika*, 2(1), 23–33.

Gallagher, C. A., Monroe, H. M., & Fish, J. L. (2000). An Iterative Approach to Classification Analysis. *Journal of Applied Statistics*, 29, 256-266.

Hamidah, I. (2013). *Aplikasi Data Mining untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Jurusan Teknik Komputer-UNIKOM).* (Tesis). UNIKOM: JBPTUNIKOMPP.

Han, J. & Kamber, M. (2006). *Data Mining Concepts and Techniques Second Edition*. San Francisco: Morgan Kaufmann.

Kunto, Y. S., & Hasana, S. N. (2006). Analisis CHAID Sebagai Alat Bantu Statistika Untuk Segmentasi Pasar. *Jurnal Manajemen Pemasaran*, 1(2), 88-98.

Lehmann, T., & Eherler, D. (2001). Responder Profiling with CHAID and Dependency Analysis. *In Data Mining for Marketing Applications Workshop*. Germany: University of Freiburg.

Miftahuddin. (2012). Penggunaan Metode CHAID (Chi Square- Automatic Interaction Detection) Pada Pohon Klasifikasi Menggunakan Satu Peubah Respon Dengan Perbandingan Taraf Nyata. *Jurnal matematika, statistika dan komputasi*, 9(1), 11-22.

Rizki, M., Umam, M. I. H., Hamzah, M. L. (2020). Aplikasi Data Mining Dengan Metode CHAID Dalam Menentukan Status Kredit. *Jurnal Sains, Teknologi dan Industri*. 18(1), 29-33.

Santoso, Singgih. (2012). *Panduan Lengkap SPSS Versi 20*. Jakarta: PT Elex Media Komputindo.

Silitonga, P. (2016). Analisis Pola Penyebaran Penyakit Pasien Pengguna Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan dengan Menggunakan Metode DBSCAN Clustering. *Jurnal Times.* 5(1), 36–39.

Sugiyono. (2015). *Statistik Nonparametris untuk Penelitian*. Bandung: Alfabeta.

Sujana, D. & Trisyan, Y. (2023). Pengkajian Resep Berdasarkan Aspek Administratif Pada Pasien Hipertensi di Puskesmas Pembangunan Garut. *Jurnal Medika Farmaka*, 1(1), 54-66.

Suyanto. (2017). *Data Mining: Untuk Klasifikasi dan Klastering Data*. Bandung: Informatika Bandung.

Tyas, A. E., Ispriyanti, D., & Sudarno. (2015). Ketepatan Klasifikasi Status Kerja di Kota Tegal Menggunakan Algoritma C4.5 dan Fuzzy K-Nearest Neighbor in Every Class (FK-NNC). *Jurnal Gaussian*, 4(4), 735-744.

Walpole, R. E. (1995). *Pengantar Statistika*. Jakarta: PT Gramedia Pustaka Utama.

Witten, Ian H, Frank, Eibe, & Hal, M.A. (2011). *Data Mining: Pratical Machine Learning Tools and Techniques, Third Edition.* Burlington: Morgan Kaufmann Publishers.

Wulandary, A. (2014). *Klasifikasi keputusan nasabah untuk menggunakan ATM dengan metode Chi-Square Automatic Interaction Detection (CHAID)*. (Skripsi). Universitas Pendidikan Indonesia: Bandung.