Analisa Keterkaitan *Risk Factor Stroke* dengan Jenis *Stroke* yang Diderita Menggunakan Algoritma ECLAT

Rio Fernando, Lia Anggraini, Alwis Nazir
Teknik Informatika, Fakultas Sains & Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau
Jl. H.R. Soebrantas Km. 15 Panam Pekanbaru, Telp. 0761-8359937
rio.fernando@students.uin-suska.ac.id

Abstrak

Stroke adalah salah satu penyakit yang sangat mematikan, namun kasus kematian yang disebabkan oleh stroke dapat diperkecil apabila kita mengetahui keterkaitan antara risk factor stroke dengan jenis stroke yang akan diderita. Penelitian ini menggunakan algoritma ECLAT untuk menganalisa keterkaitan antara risk factor stroke dengan jenis stroke yang akan diderita pasien. Data yang digunakan berdasarkan data rekam medis pasien stroke di RSUD Puri Husada Tembilahan dari tahun 2011 hingga tahun 2015 dengan total record sebanyak 700 record. Hasil analisa keterkaitan yang didapatkan dengan menggunakan tools R menghasilkan beberapa rule utama dengan nilai support tertinggi sebesar 41% untuk diagnosa stroke iskemik dan 14% untuk diagnosa stroke hemoragik. Tingkat akurasi hasil analisa tersebut menghasilkan nilai akurasi tertinggi sebesar 97.23% dan terendah sebesar 6.66%.

Kata kunci: *stroke, aturan asosiasi, data mining, ECLAT*

Abstract

Stroke is an extremely deadly disease but cases of death caused by stroke can be minimized if we know the relation between stroke risk factor with type of stroke that will be suffered. This study used ECLAT Algorithm to analyze the relations between stroke's risk factor with the type of stroke that will be suffered by patients. We used medical records of stroke patients from 2011 to 2015 in Puri Husada Tembilahan Hospital with a total number of 700 records. By using R as tools, the results of the analysis produced some representative rules with the highest value of support at 41% for ischemic stroke and 14% for hemorrhagic stroke. The representative rule analysis resulted the highest accuracy value at 95.23% and the lowest accuracy value at 6.66%.

Keywords*: stroke, association rule, data mining, ECLAT*

1. Pendahuluan

Pada negara-negara berkembang, penyakit *stroke* merupakan penyebab kematian paling umum mengungguli kanker sebagai penyebab kematian paling umum kedua [1]. Berdasarkan hasil survei Badan Penelitian dan Pengembangan Kesehatan (Balitbangkes) Kementerian Kesehatan tahun 2014, *stroke* merupakan penyebab kematian utama di Indonesia. Hal ini didukung oleh data diagnosis Nakes/ gejala yang memperkirakan jumlah penderita *stroke* di Indonesia mencapai 2.137.941 orang (12,1%). Maka dapat disimpulkan bahwa penelitian mengenai penyakit *stroke* masih sangat diperlukan sebagai kontribusi untuk mengurangi tingginya jumlah kematian akibat stroke di Indonesia.

American Heart/Stroke Association (AHA/ASA) menyimpulkan bahwa identifikasi awal mengenai berbagai faktor tertentu yang meningkatkan risiko terjadinya *stroke* dapat mengurangi risiko *stroke* itu sendiri [2]. Penelitian mengenai identifikasi faktor *stroke* telah dilakukan sebelumnya [3] yang mengkombinasikan identifikasi secara fisik, radiologik, dan labratorik untuk menghasilkan *screening* cepat mengenai faktor yang mempengaruhi risiko terjadinya *stroke*. Hasil penelitian tersebut memiliki tingkat keakuratan paling rendah 67% dan paling tinggi 74%. Penelitian lainnya yang lebih mengandalkan sisi teknologi informasi [4] menganalisa berbagai faktor mengenai *stroke iskemik* dengan menggunakan beberapa metode *data mining* yang berbeda, antara lain; *support vector machine* (SVM), *stochastic gradient boosting* (SGB) dan *penalized logistic regression* (PLR). Penelitian ini menghasilkan tingkat keakuratan yang cukup stabil (0.9789 (0.9470-0.9942) SVM, 0.9737 (0.9397-0.9914) SGB, dan 0.8947 (0.8421-0.9345) PLR berdasarkan CI 95%.

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

Penelitian yang telah dilakukan memiliki hasil yang bermanfaat dalam mengidentifikasi berbagai faktor penyebab penyakit *stroke*, namun hasil penelitian tersebut hanya berfokus pada faktor *stroke iskemik*. Penelitian yang sekaligus berfokus pada jenis *stroke* lainnya seperti *stroke hemoragik* masih sangat jarang ditemukan, sehingga sebuah pendekatan baru sangat diperlukan untuk menganalisa dan mengidentifikasi berbagai faktor penyebab *stroke*, terutama penelitian yang bisa mengidentifikasi kaitan berbagai faktor dengan berbagai jenis penyakit *stroke*. Penelitian yang kami lakukan merupakan suatu inovasi baru karena dapat mengidentifikasi kaitan antara berbagai faktor penyebab *stroke* dengan jenis penyakit *stroke* yang ada.

Pada penelitian ini, kami menggunakan algoritma ECLATsebagai metode dan R sebagai *tools* untuk menganalisa kaitan antara berbagai faktor penyebab *stroke* dengan jenis penyakit *stroke* yang ada. Tujuan dari penelitian adalah mencari dan menghasilkan *rule* sekaligus kesimpulan, mengenai kaitan antara berbagai faktor penyebab *stroke* dan berbagai jenis penyakit *stroke*. Diharapkan dari penelitian ini akan memiliki imbas yang cukup besar pada komunitas medis, salah satunya dengan penurunan jumlah kematian akibat *stroke*.

2. Metode Penelitian

2.1. Pengumpulan Data

Sumber data yang digunakan merupakan data primer pasien penyakit *stroke* dengan kasus *stroke iskemik* dan *stroke hemoragik* yang diperoleh dari Rumah Sakit Umum Daerah (RSUD) Puri Husada Tembilahan. Data ini merupakan kumpulan data dari tahun 2011 hingga tahun 2015 yang berjumlah mencapai 700 data. Data rekam medis pasien *stroke* ini terdiri dari 10 atribut.

Tabel 1. Atribut data rekam medis pasien *stroke*

Atribut	Keterangan	*Value*
Rekam Medik	ID rekam medik pasien	*String*
Usia	Umur pasien	*Numeric*
Gender	Jenis kelamin pasien	*String*
Sistol	Kadar sistol pasien	*Numeric*
Diastol	Kadar diastol pasien	*Numeric*
HDL	Kadar HDL pasien	*Numeric*
LDL	Kadar LDL pasien	*Numeric*
Kolesterol Total	Kolesterol Total pasien	*Numeric*
TGA	Kadar triglycerida pasien	*Numeric*
Diagnosa	Jenis stroke pasien	*String*

Tabel 2. Data *Sampling*

No	MR	Usia	*Gender*	Sistol	Diasto	HDL	LDL	Kolesterol Total	TGA	Diagnosa
1	00.05.67	63	P	190	100	44	87	143	60	Stroke Iskemik
2	00.06.60	70	L	140	90	70	129	223	116	Stroke Hemoragik
3	00.10.56	46	L	180	100	67	122	201	76	Stroke Hemoragik
4	00.10.78	51	L	120	80	61	68	142	62	Stroke Iskemik
...										
630	25.99.46	46	L	150	140	47	70	160	215	Stroke Iskemik

Untuk pembagian data yang akan digunakan untuk data *sampling* dan data *testing* dibagi secara random menjadi 90% data *sampling* dan 10% data *testing*.

Tabel 3. Rincian Pembagian Data

No	Jenis Data	Jumlah
1	Data *Sampling*	630 (90%)
2	Data *Testing*	70 (10%)
	Total	700 (100%

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

2.2. Pengolahan Data

Setelah semua data yang dibutuhkan telah berhasil dikumpulkan, selanjutnya adalah penerapan proses pengolahan data ini mengikuti Tahapan KDD sehingga data bisa diproses dalam *data mining*.

A. Data Selection

Data selection atau proses seleksi data merupakan proses untuk memilih atribut-atribut mana saja yang digunakan untuk penelitian dan membuang atribut yang tidak dipakai pada penelitian. Dari 10 atribut yang terdapat pada data awal atribut, MR merupakan satu-satunya atribut yang bukan merupakan *risk factor stroke*, sehingga atribut MR dieliminasi dan menyisakan 9 atribut yaitu usia, *gender*, HDL, LDL, sistol, diastol, kolesterol total TGA dan diagnosa.

B. Data Processing

Data Preprocessing merupakan tahap pembersihan data. Pada tahap ini akan dikoreksi data yang memiliki duplikasi, inkonsistensi dan *missing value* pada data rekam medis pasien *stroke*. Terdapat 0 data yang memiliki duplikasi, 17 data yang tidak konsisten dan 5 data yang hilang atau missing value.

Tabel 4. Contoh Inkonsistensi Data Rekam Medis Pasien *Stroke*

No	Usia	*Gender*	Sistol	Diastol	HDL	LDL	Kolesterol Total	TGA	Diagnosa
98	46	L	170	110	50	211	291	151	Stroke Infark
99	48	P	200	90	42	132	244	347	Stroke Iskemik
100	75	L	180	90	50	63	150	185	Stroke Iskemik
…									
105	71	P	200	110	42	106	165	82	Stroke Pendarahan

Pada baris ke 98 tertulis diagnosa berupa *stroke infark* dan pada baris ke 105 tertulis diagnosa berupa *stroke* pendarahan. *Stroke infark* merupakan istilah lain dari *stroke iskemik* sedangkan *stroke* pendaharan merupakan istilah lain dari *stroke hemoragik*. Dalam perbaikan inkonsistensi data ini setiap diagnosa yang tertulis sebagai *stroke infark* akan diperbaiki menjadi *stroke iskemik*, dan setiap diagnosa yang tertulis sebagai *stroke* pendarahan akan diperbaiki menjadi *stroke hemoragik*.

Tabel 5. Contoh *Missing Value* Data Rekam Medis Pasien *Stroke*

No	Usia	*Gender*	Sistol	Diastol	HDL	LDL	Kolesterol Total	TGA	Diagnosa
1	63	P	190	100	44	87	143	60	Stroke Iskemik
..									
49	76	P	180	100	45	196	264	115	Stroke Iskemik
50	53	L	170	100	44	135	201	?	Stroke Iskemik
51	59	L	190	100	45	181	247	105	Stroke Hemoragik
..									
118	55	L	140	90	32	120	183	514	Stroke Iskemik
119	68	L	?	?	66	102	188	97	Stroke Iskemik

Pada Tabel 5 terdapat data yang memiliki *missing value* yakni pada baris ke 50 dan baris ke 119. Terdapat beberapa metode dalam mengatasi *missing value* pada data dengan nilai *numeric*, salah satunya adalah dengan mengisi data kosong dengan nilai rata-rata dari dua baris tetangga yang terdekat pada kolom yang sama. Berikut ini cara untuk mencari nilai rata-rata ($\bar{X}$) dari nilai data baris terdekat pertama (Xi) dan baris terdekat kedua (Xj) [23].

1) $\bar{X}50 = \frac{x_i + x_j}{2} = \frac{115+105}{2} = 110$
2) $\bar{X}119 = \frac{x_i + x_j}{2} = \frac{140+120}{2} = 130$
3) $\bar{X}119 = \frac{x_i + x_j}{2} = \frac{90+90}{2} = 90$

Pada kasus *missing value* pertama ($\bar{X}50$) adalah data pada kolom ke 50 sehingga (Xi) adalah data pada kolom ke 49, dan (Xj) adalah data pada kolom ke 51. Sedangkan untuk kasus *missing value* kedua dan ketiga ($\bar{X}119$) adalah data pada kolom ke 119 sehingga (Xi) adalah

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

data pada kolom ke 118, dan (Xj) adalah data pada kolom ke 120. Pengisian *missing value* pada data dengan menggunakan metode pencarian rata-rata dipilih agar persebaran data antara baris yang berdekatan tidak terlalu jauh dan lebih konsisten.

C. Data Transformation

Proses transformasi data yang dilakukan adalah normalisasi data dimana atribut data dibuat dalam skala tertentu agar menjadi kisaran data yang lebih kecil sehingga persebaran datanya tidak terlalu jauh. Atribut data yang mengalami normalisasi adalah usia yang dikelompokkan [2] menjadi tiga bagian (<65, 65-75, >75), diikuti oleh pengelompokan nilai *risk factor* lainnya [5] dengan detail sebagai berikut; sistol (<90, 90-120, >120), diastol (<60, 60-80, >80), HDL (<40, 40-60, >60), LDL (<100, 100-129, 130-159, 160-190, >190), kolesterol total (<200, 200-240, >240), dan TGA (<150, 150-199, 200-500, >500).

Tabel 6. Hasil *Data Transformation*

No	Usia	*Gender*	Sistol	Diastol	HDL	LDL	Kolesterol Total	TGA	Diagnosa
1	<65	P	>120	>80	40-60	<100	<200	<150	Stroke Iskemik
2	65-75	L	>120	>80	>60	100-129	200-240	<150	Stroke Hemoragik
3	<65	L	>120	>80	>60	100-129	200-240	<150	Stroke Hemoragik
4	<65	L	90-120	60-80	>60	<100	<200	<150	Stroke Iskemik
…									
630	<65	L	>120	>80	40-60	<100	<200	200-500	Stroke Iskemik

3. Hasil dan Pembahasan

Pada pembahasan ini dilakukan perbandingan konsistensi hasil analisa algoritma ECLAT secara manual terhadap hasil analisa ECLAT menggunakan *tools* R. Perbandingan konsistensi dilakukan dengan menggunakan sebagian dari data *sampling* yang telah dikonversi ke bentuk data vertikal (Tabel 7).

Kemudian terhadap data vertikal yang ada dilakukan penyilangan 2-*itemsets* antara transaksi yang mengandung nilai Diagnosa=*Stroke Iskemik* dan Diagnosa= *Stroke Hemoragik* dengan transaksi lainnya. Hasil penyilangan tersebut menghasilkan analisa berupa kumpulan *itemset*, kumpulan *itemset* inilah yang kemudian akan digunakan sebagai *rule* hasil analisa.

Setelah *rule* didapatkan maka selanjutnya dilakukan pencarian nilai *support* dan nilai *confidence* dari *rule* terkait. Nilai *support* adalah nilai perbandingan antara transaksi yang mengandung *rule* yang dihasilkan dengan jumlah total transaksi. Menggunakan Persamaan *support* dan Persamaan *confidence* maka didapatkan nilai *support* dan nilai *confidence* untuk setiap *rule* (Tabel 8). Hasil analisa dengan penerapan metode ECLATsecara manual tersebut menghasilkan hasil yang persis sama dengan hasil analisa menggunakan tools R (Gambar 1).

Tabel 7. Data Vertikal

Transaksi Vertikal	
Itemset	**TID List**
Usia=<65	(1,3,4,5)
Usia=65-75	(2)
Gender=L	(2,3,4,5)
Gender=P	(1)
Sistol=90-120	(4)
Sistol=>120	(1,2,3,5)
Diastol=60-80	(4)
Diastol=>80	(1,2,3,5)
HDL=40-60	(1)
HDL=>60	(2,3,4,5)
LDL=<100	(1,4)
LDL=100-129	(2,3)
LDL=160-190	(5)
Kolesterol.Total=<200	(1,4)
Kolesterol.Total=200-240	(2,3)
Kolesterol.Total=>240	(5)
TGA=<150	(1,2,3,4,5)
Diagnosa=Stroke Iskemik	(1,4,5)
Diagnosa=Stroke Hemoragik	(2,3)

Min Support = 2	
Itemset	**TID List**
Usia=<65	(1,3,4,5)
Gender=L	(2,3,4,5)
Sistol=>120	(1,2,3,5)
Diastol=>80	(1,2,3,5)
HDL=>60	(2,3,4,5)
LDL=<100	(1,4)
LDL=100-129	(2,3)
Kolesterol.Total=<200	(1,4)
Kolesterol.Total=200-240	(2,3)
TGA=<150	(1,2,3,4,5)
Diagnosa=Stroke Iskemik	(1,4,5)
Diagnosa=Stroke Hemoragik	(2,3)

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

Tabel 8. Hasil Analisa Manual

Itemset	Nilai *Support*	Nilai *Confidence*
(Diagnosa=Stroke Iskemik, Usia=<65)	0,60	1,00
(Diagnosa=Stroke Iskemik, Gender=L)	0,40	0,67
(Diagnosa=Stroke Iskemik, Sistol=>120)	0,40	0,67
(Diagnosa=Stroke Iskemik, Diastol=>80)	0,40	0,67
(Diagnosa=Stroke Iskemik, HDL=>60)	0,40	0,67
(Diagnosa=Stroke Iskemik, LDL=<100)	0,40	0,67
(Diagnosa=Stroke Iskemik, Kolesterol.Total=<200)	0,40	0,67
(Diagnosa=Stroke Iskemik, TGA=<150)	0,40	1,00
(Diagnosa=Stroke Hemoragik, Gender=L)	0,40	1,00
(Diagnosa=Stroke Hemoragik, Sistol=>120)	0,40	1,00
(Diagnosa=Stroke Hemoragik, Diastol=>80)	0,40	1,00
(Diagnosa=Stroke Hemoragik, HDL=>60)	0,40	1,00
(Diagnosa=Stroke Hemoragik, LDL=100-129)	0,40	1,00
(Diagnosa=Stroke Hemoragik, Kolesterol.Total=200-240)	0,40	1,00
(Diagnosa=Stroke Hemoragik, TGA=<150)	0,40	1,00

```
> inspect(itemsets)
   items                                                    support
1  {TGA=<150,Diagnosa=Stroke Hemoragik}                     0.4
2  {Sistol=>120,Diagnosa=Stroke Hemoragik}                  0.4
3  {Diastol=>80,Diagnosa=Stroke Hemoragik}                  0.4
4  {Gender=L,Diagnosa=Stroke Hemoragik}                     0.4
5  {HDL=>60,Diagnosa=Stroke Hemoragik}                      0.4
6  {LDL=100-129,Diagnosa=Stroke Hemoragik}                  0.4
7  {Kolesterol.Total=200-240,Diagnosa=Stroke Hemoragik}     0.4
8  {Kolesterol.Total=<200,Diagnosa=Stroke Iskemik}          0.4
9  {LDL=<100,Diagnosa=Stroke Iskemik}                       0.4
10 {TGA=<150,Diagnosa=Stroke Iskemik}                       0.6
11 {Sistol=>120,Diagnosa=Stroke Iskemik}                    0.4
12 {Diastol=>80,Diagnosa=Stroke Iskemik}                    0.4
13 {Usia=<65,Diagnosa=Stroke Iskemik}                       0.6
14 {Gender=L,Diagnosa=Stroke Iskemik}                       0.4
15 {HDL=>60,Diagnosa=Stroke Iskemik}                        0.4
```

Gambar 1. Hasil Analisa Menggunakan *Tools* R

Setelah melakukan perbandingan konsistensi antara analisa ECLATsecara manual dengan analisa ECLAT menggunakan *tools* R, selanjutnya dilakukan proses analisa terhadap data secara keseluruhan. Dalam analisa menggunakan *tools* R ini nilai *minimum length* (*minlen*) dan *maximum length (maxlen)* yang digunakan adalah 4 (dimana setiap keterkaitan data yang memiliki jumlah atribut kurang atau lebih dari 4 tidak akan ditampilkan) dengan nilai *support itemset* terkecil 0.1 (dimana *itemset* yang memiliki nilai *support* lebih kecil dari 10% tidak akan ditampilkan). Hasil analisa tersebut menghasilkan 109 *rule* (Tabel 9) dengan diagnosa *stroke iskemik* dan 10 *rule* dengan diagnosa *stroke hemoragik* (Tabel 10).

Tabel 9. Hasil Analisa Stroke Iskemik

No	*Risk Factor 1*	*Risk Factor 2*	*Risk Factor 3*	Diagnosa	Nilai *Support*
1	Gender=L	Sistol=>120	Diastol=>80	Diagnosa=Stroke Iskemik	0.4158730
2	Sistol=>120	Diastol=>80	TGA=<150	Diagnosa=Stroke Iskemik	0.4095238
3	Usia=<65	Sistol=>120	Diastol=>80	Diagnosa=Stroke Iskemik	0.3634921
4	Gender=L	Sistol=>120	TGA=<150	Diagnosa=Stroke Iskemik	0.3269841
5	Sistol=>120	Diastol=>80	HDL=40-60	Diagnosa=Stroke Iskemik	0.3126984
6	Gender=L	Diastol=>80	TGA=<150	Diagnosa=Stroke Iskemik	0.3047619
7	Sistol=>120	Diastol=>80	Kolesterol.Total=<200	Diagnosa=Stroke Iskemik	0.3015873
8	Sistol=>120	Kolesterol.Total=<200	TGA=<150	Diagnosa=Stroke Iskemik	0.2682540
9	Usia=<65	Gender=L	Sistol=>120	Diagnosa=Stroke Iskemik	0.2666667
...	Usia=<65	Sistol=>120	TGA=<150	Diagnosa=Stroke Iskemik	0.2634921
109	Gender=L	LDL=<100	TGA=<150	Diagnosa=Stroke Iskemik	0.1000000

Tabel 10. Hasil Analisa Stroke Hemoragik

No	Risk Factor 1	Risk Factor 2	Risk Factor 3	Diagnosa	Support
1	Usia=<65	Sistol=>120	Diastol=>80	Stroke Hemoragik	0.1396825
2	Sistol=>120	Diastol=>80	TGA=<150	Stroke Hemoragik	0.1349206
3	Gender=L	Sistol=>120	Diastol=>80	Stroke Hemoragik	0.1253968
4	Usia=<65	Sistol=>120	TGA=<150	Stroke Hemoragik	0.1063492
5	Sistol=>120	Diastol=>80	Kolesterol.Total=<200	Stroke Hemoragik	0.1063492
6	Gender=L	Sistol=>120	TGA=<150	Stroke Hemoragik	0.1031746
7	Usia=<65	Diastol=>80	TGA=<150	Stroke Hemoragik	0.1031746
8	Usia=<65	Gender=L	Sistol=>120	Stroke Hemoragik	0.1031746
9	Usia=<65	Gender=L	Diastol=>80	Stroke Hemoragik	0.1000000
10	Sistol=>120	Diastol=>80	HDL=40-60	Stroke Hemoragik	0.1000000

Setelah dilakukan analisa dan ditemukannya *rule* keterkaitan antar atribut data, maka dilakukan pengujian terhadap hasil analisa tersebut. Tahap pengujian diperlukan sebagai acuan bahwa *rule* keterkaitan yang dihasilkan memiliki tingkat akurasi yang layak untuk diimplementasikan.

3.1. Penyeleksian Representative Rule dengan menggunakan metode Quota Sampling

Analisa mengenai keterkaitan antara *risk factor stroke* dengan jenis *stroke* yang diderita menghasilkan 119 *rule* dengan dua jenis diagnosa yakni 109 *rule* dengan diagnosa Stroke Iskemik dan dan 10 *rule* dengan diagnosa Stroke Hemoragik. Untuk menerapkan metode *quota sampling* maka ditetapkan bahwa kriteria *representative rule* yang akan digunakan adalah sebagai berikut [15]:

1. Mewakili sedikitnya 5% dari total *rule*, dengan jumlah total 119 *rule* maka *representative rule* yang digunakan berjumlah paling sedikit sebanyak 6 *rule*.
2. Diseleksi secara *descending* dimulai dari nilai *support* yang tertinggi.
3. *Rule* ke *n* harus memiliki diagnosa yang berbeda dengan *rule* ke *n*-1.
4. *Rule* ke *n*+x harus memiliki *risk factor* kolektif yang berbeda dengan *rule* ke 1 hingga *rule* ke x-1.

Dengan menerapkan kriteria di atas maka *rule* yang memenuhi syarat kriteria *representative rule* dapat dilihat pada Tabel 11.

Tabel 11. Daftar *Representative Rule*

No	Risk Factor 1	Risk Factor 2	Risk Factor 3	Diagnosa	Support
1	Gender=L	Sistol=>120	Diastol=>80	Stroke Iskemik	0.4158730
2	Usia=<65	Sistol=>120	Diastol=>80	Stroke Hemoragik	0.1396825
3	Sistol=>120	Diastol=>80	TGA=<150	Stroke Iskemik	0.4095238
4	Usia=<65	Sistol=>120	TGA=<150	Stroke Hemoragik	0.1063492
5	Gender=L	Sistol=>120	TGA=<150	Stroke Iskemik	0.3269841
6	Sistol=>120	Diastol=>80	Kolesterol Total=<200	Stroke Hemoragik	0.1063492

Representative rule di atas dapat diterjemahkan ke bentuk *narative* sebagai berikut:

A. Jika *gender* laki laki dengan sistol di atas 120 dan diastol di atas 80 maka dengan nilai *support* 41%, stroke yang akan diderita adalah *stroke iskemik*.
B. Jika usia di bawah 65 tahun dengan sistol di atas 120 dan diastol di atas 80 maka dengan nilai *support* 14% stroke yang akan diderita adalah *stroke hemoragik*.
C. Jika sistol di atas 120 dengan diastol di atas 80 dan TGA di bawah 150 maka dengan nilai *support* 40% stroke yang akan diderita adalah *stroke iskemik*.
D. Jika usia di bawah 65 tahun dengan sistol di atas 120 dan TGA di bawah 150 maka dengan nilai *support* 10% stroke yang akan diderita adalah *stroke hemoragik*.
E. Jika gender laki laki dengan sistol di atas 120 dan TGA di bawah 150 maka dengan nilai *support* 32% stroke yang akan diderita adalah *stroke iskemik*.
F. Jika sistol di atas 120 dengan diastol di atas 80 dan kolesterol total di bawah 200 maka dengan nilai *support* 10% stroke yang akan diderita adalah *stroke hemoragik*.

3.2. Pengujian akurasi dengan menggunakan metode *independent t-Test*

Setelah beberapa *representative rule* didapatkan, maka selanjutnya dilakukan perbandingan *representative rule* dengan data *testing*. Pada Tabel 12 dapat dilihat *preview* data *testing* yang telah dipersiapkan.

Tabel 12. *Preview* Data *Testing*

No	Usia	*Gender*	Sistol	Diastol	HDL	LDL	Kolesterol Total	TGA	Diagnosa
1	>75	P	>120	>80	40-60	160-190	>240	200-500	Stroke Iskemik
2	65-75	L	>120	>80	40-60	100-129	200-240	150-199	Stroke Iskemik
3	65-75	P	>120	>80	>60	100-129	<200	<150	Stroke Hemoragik
4	>75	L	>120	>80	<40	>190	>240	<150	Stroke Iskemik
5	>75	L	>120	>80	<40	>190	>240	<150	Stroke Iskemik
6	65-75	P	>120	>80	40-60	<100	200-240	200-500	Stroke Iskemik
...									
70	65-75	L	>120	>80	40-60	<100	<200	150-199	Stroke Iskemik

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

Berikut adalah hasil tes akurasi dari keseluruhan *representative rule* yang ada:

A. Rule 1: Jika *gende*r laki laki dengan sistol di atas 120 dan diastol di atas 80 maka dengan nilai *support* 41%, stroke yang akan diderita adalah *stroke iskemik*. Hasil pengujian ke data *testing* menghasilkan tingkat akurasi sebesar 93.75%.

B. Rule 2: Jika usia di bawah 65 tahun dengan sistol di atas 120 dan diastol di atas 80 maka dengan nilai *support* 14% stroke yang akan diderita adalah *stroke hemoragik*. Hasil pengujian ke data *testing* menghasilkan tingkat akurasi sebesar 13.04%.

C. Rule 3: Jika sistol di atas 120 dengan diastol di atas 80 dan TGA di bawah 150 maka dengan nilai *support* 40% stroke yang akan diderita adalah *stroke iskemik*. Hasil pengujian ke data *testing* menghasilkan tingkat akurasi sebesar 76.47%.

D. Rule 4: Jika usia di bawah 65 tahun dengan sistol di atas 120 dan TGA di bawah 150 maka dengan nilai *support* 10% stroke yang akan diderita adalah *stroke hemoragik*. Hasil pengujian ke data *testing* menghasilkan tingkat akurasi sebesar 6.66%.

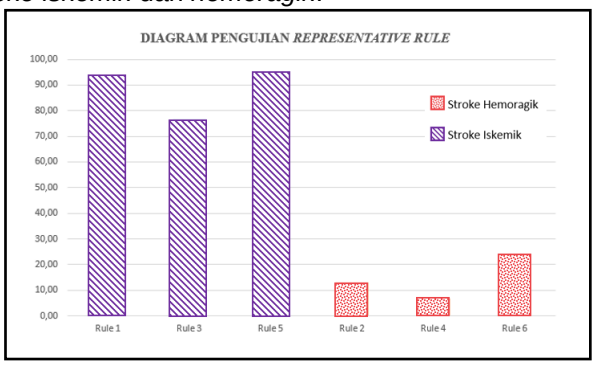
E. Rule 5: Jika gender laki laki dengan sistol di atas 120 dan TGA di bawah 150 maka dengan nilai *support* 32% stroke yang akan diderita adalah *stroke iskemik*. Hasil pengujian ke data *testing* menghasilkan tingkat akurasi sebesar 95.23%.

F. Rule 6: Jika sistol di atas 120 dengan diastol di atas 80 dan kolesterol total di bawah 200 maka dengan nilai *support* 10% stroke yang akan diderita adalah *stroke hemoragik*. Hasil pengujian ke data *testing* menghasilkan tingkat akurasi sebesar 24%.

3.3. Kesimpulan Pengujian

Berdasarkan hasil pengujian di atas maka diperoleh beberapa kesimpulan:

1. Kadar sistol yang lebih dari 120 mmHg merupakan *risk factor* yang sangat signifikan terhadap terjadinya *stroke*, hal ini terbukti dengan kemunculan *risk factor* Sistol=>120 di setiap *representative rule* yang di dapat.
2. *Gender* laki-laki memiliki risiko menderita *stroke* lebih tinggi dari pada *gender* perempuan. Hal ini diperlihatkan dengan munculnya *risk factor* Gender=L sebanyak 33% dari total *representative rule* sedangkan Gender=P tidak muncul sama sekali sebagai *risk factor* pada *representative rule*.
3. *Representative rule* dengan diagnosa *stroke iskemik* memiliki tingkat akurasi yang lebih tinggi daripada *representative rule* dengan diagnosa *stroke hemoragik*. Hal ini diperlihatkan dari hasil pengujian dimana *representative rule* dengan diagnosa *stroke iskemik* memiliki akurasi terendah 76.47% dan akurasi tertinggi 95.23% sedangkan *representative rule* dengan diagnosa *stroke hemoragik* memiliki tingkat akurasi terendah 6.66% dan akurasi tertinggi 24%. Pada Gambar 2 merupakan diagram hasil pengujian *representative rule* dengan diagnosa *stroke iskemik* dan *hemoragik*.



Gambar 2. Diagram Penggujian *Representative Rule*

Hal ini dikarenakan *representative rule* dengan diagnosa *stroke hemoragik* memiliki nilai *support* yang lebih kecil daripada *representative rule* dengan diagnosa *stroke iskemik*, dimana *representative rule* dengan diagnosa *stroke hemoragik* memiliki nilai *support* terendah 0.1063492 dan nilai *support* tertinggi 0.1396825 sedangkan *representative rule* dengan diagnosa *stroke iskemik* memiliki nilai *support* terendah 0.3269841 dan nilai *support* tertinggi 0.4158730.

4. Kesimpulan

Penelitian yang kami lakukan mampu mengidentifikasi keterkaitan antara *risk factor stroke* dengan jenis *stroke* yang diderita dalam bentuk *representative rule* yang didapat dari analisa menggunakan data sampling sebesar 630 data. Sedangkan tingkat akurasi dari hasil analisa *representative rule* yang dihitung dengan menggunakan metode independent t-Test dengan menggunakan 70 data, mencapai tingkat keakuratan tertinggi 95.23% dan terendah 6.66%. Untuk penelitian selanjutnya, *risk factor* yang digunakan hendaknya lebih lengkap sesuai dengan *risk factor* keseluruhan *stroke*, termasuk di antaranya ras, faktor genetik, dan tingkat aktifitas fisik.

Daftar Pustaka

[1] Yigit, M. O. The relationship between anemia and recurrence of ischemic stroke in patients with Trousseau's syndrome: A retrospective cross-sectional study.*Turkish Journal of Emergency Medicine*. 2016.
[2] Meschia, J. F., & Bushnell, C. M.-C. Guidelines for the Primary Prevention of Stroke. A Statement for Healthcare Professionals From the American. 2014.
[3] Ohara, T. M. Rapid Identification of Type A Aortic Dissection as a Cause of Acute Ischemic Stroke. *Journal of Stroke and Cerebrovascular Diseases*. 2016.
[4] Arslan, A. K. *Different medical data mining approaches based prediction of ischemic stroke. Computer Methods and Programs in Biomedicine.* 2016.
[5] Hao, W. A. The LDL-HDL Profile Determines the Risk of Atherosclerosis: A Mathematical Model. *Mathematical Biosciences Institute and the National Science Foundation Journal*. 2014.
[6] Al Essa, A. R. Data Mining and Warehousing. *American Society for Engineering Education (ASEE Zone 1) Journal*. 2014
[7] Angueraa, A. J. Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry. *Computational and Structural Biotechnology Journal*. 2016.
[8] de Rada, D., & Martín, M. Random Route and Quota Sampling: Do They Offer Any Advantage over Probably Sampling Methods?. *Open Journal of Statistics*. 2014.
[9] de Winter, J. Using the Student's t-test with extremely small sample sizes. Practical Assessment, Research & Evaluation: *A peer-reviewed electronic journal*. 2013.
[10] Gupta, D. A. Mining Association Rules from Infrequent Itemsets: A Survey. *International Journal of Innovative Research in Science, Engineering and Technology*. 2013.
[11] Hatano, S. A. *Cerebrovascular disease in the community: results of a WHO collaborative study*. Bull World Health Organ. 1980.
[12] Ihaka, R., & Gentleman, R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*. 1996.
[13] Jellinger, P. S. American Association of Clinical Endocrinologists' Guidelines for Management of Dyslipidemia and Prevention of Atherosclerosis. *AACE Lipid and Atherosclerosis Guidelines, Endocr Practice*. 2012; 18.
[14] Kaur, M. U. ECLAT Algorithm for Frequent Itemsets Generation. *International Journal of Computer Systems*. 2014.
[15] Kish, L. On Quota Sampling. *Working Paper, Universidad de Michigan*. 1998.
[16] Ramaraj, E., & Venkatesan, N. An Efficient Pattern Mining Analysis in Health Care Database. *Journal of Theoretical & Applied Information Technology*. 2009.
[18] Rothman, J., & Dawn, M. Statisticians Can Be Creative Too. *Journal of the Market Research Society*. 1989.
[19] Saxena, A. S. A Survey on frequent pattern mining methods - Apriori, Eclat, FP growth. *International Journal of Engineering Development and Research*. 2014.
[20] Schmidt-Thieme, L. Algorithmic Features of Eclat. *University of Freiburg's Institute for Computer Science's Journal*. 2003.
[21] Syvajarvi, A. Data Mining in Public and Private Sectors: Organizational and Government Applications. New York: Hersey. 2010.
[22] United Nations' Statistical Office. Provisional Guidelines on Standard International Age Classifications. New York: United Nations. 1982.
[23] Vijayarani, D. S. An Efficient Algorithm for Mining Frequent Items in Data Streams. *International Journal of Innovative Research in Computer and Communication Engineering*. 2013.
[24] Zhao, Y. R and Data Mining: Examples and Case Studies. Elsevier. 2013.