Perbandingan Klasifikasi Metode *Naive* Bayes dan Metode *Decision Tree* Algoritma (J48) pada Pasien Penderita Penyakit *Stroke* di RSUD Abdul Wahab Sjahranie Samarinda

Comparison of the Classification for Naive Bayes *Method and the Decision Tree Algorithm* (*J48*) *for Stroke Patients in* Abdul Wahab Sjahranie Samarinda *Hospital*

Irene Lishania[1], Rito Goejantoro[2], dan Yuki Novia Nasution[3]
[1,2]Laboratorium Statistika Komputasi FMIPA Universitas Mulawarman
[3]Laboratorium Matematika Komputasi FMIPA Universitas Mulawarman
E-mail: irenelishania@gmail.com

Abstract

Classification is a technique to form a model of the data that has not been classified, then the model can be used to classify new data. Naive Bayes *is a classification using probability method based on the* Bayes *theorem with a strong assumption of independence. The decision tree algorithm* (*J48*) *is an implementation of the algorithm* (*C4.5*) *that produces decision trees. In this research, will be compared the results of classification accuracy with the naive* Bayes *method and the decision tree algorithm* (*J48*) *in stroke patients. That is, a person who has stroke will be classified by using the data of patients in* Abdul Wahab Sjahranie Samarinda *Hospital with 7 factors, namely age, gender, blood pressure, diabetes mellitus, dyslipidemia, uric acid levels and heart disease. The results showed that the decision tree algorithm* (*J48*) *method has the higher level of accuracy than the method naive* Bayes *for stroke classification.*

Keywords: Classification, Decision Tree J48, Naive Bayes*, Stroke.*

Pendahuluan

Pertukaran informasi di zaman modern ini telah sampai pada era digital. Hal ini ditandai dengan semakin dibutuhkannya teknologi berupa komputer dan jaringan internet sebagai sarana utama penyampaian informasi. Seiring berjalannya waktu informasi yang beredar melalui komputer semakin banyak. Dampak yang disebabkan oleh arus informasi yang cepat ini adalah semakin banyaknya data-data yang tersimpan di dalam jaringan. Data yang dihasilkan oleh teknologi ini tidak hanya berguna di satu bidang saja, tetapi hampir di semua bidang kehidupan.

Untuk memanfaatkan data yang tersebar sangat banyak di dunia digital ini, diperlukan suatu alat untuk mengolah dokumen-dokumen sehingga informasinya dapat terserap dan tersajikan dengan baik. Salah satu bentuk dari pengolahan suatu data yaitu *data mining*. *Data mining* adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*.

Data mining juga merupakan proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk menguraikan dan mengidentifikasi informasi yang bermanfaat. Banyak fungsi yang dapat dilakukan menggunakan *data mining* di antaranya adalah *classification*, *clustering*, *feature selection* dan *association rule mining*. Ada beberapa macam pengklasifikasian dalam *data mining* yaitu *decision tree*, *naive* Bayes, *Support Vector Machine* (SVM) dan lain-lain (Larose,2005).

Naive Bayes merupakan pengklasifikasian dengan metode probabilitas yang ditemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes. Teorema tersebut dikombinasikan dengan *naive* di mana diasumsikan kondisi antar petunjuk (atribut) saling bebas (Bustami, 2013). Sedangkan algoritma J48 merupakan implementasi dari algoritma C4.5 yang memproduksi *decision tree*. Ini merupakan standar algoritma yang digunakan dalam *machine learning*. Salah satu pengaplikasian dari *naive* Bayes dan *decision tree* algoritma (J48) yaitu pada bidang kesehatan.

Setiap tahunnya lebih dari 36 juta orang meninggal karena Penyakit Tidak Menular (PTM) (63% dari seluruh kematian). Lebih dari 9 juta kematian yang disebabkan oleh penyakit tidak menular terjadi sebelum usia 60 tahun, dan 90% dari kematian "dini" tersebut terjadi di negara berpenghasilan rendah dan menengah. Secara global PTM penyebab kematian nomor satu setiap tahun adalah penyakit kardiovaskuler. Penyakit kardiovaskuler adalah penyakit yang disebabkan gangguan fungsi jantung dan pembuluh darah, seperti penyakit jantung koroner, penyakit gagal jantung, hipertensi dan *stroke* (Kemenkes, 2014).

Penyakit *stroke* merupakan penyakit yang menjadi penyebab kematian nomor tiga tertinggi di Indonesia setelah penyakit jantung dan kanker. Serangan *stroke* selalu datang mendadak tanpa tanda-tanda pasti. *Stroke* adalah penyakit serebrovaskuler (pembuluh darah otak) yang ditandai dengan kematian jaringan otak (*infark* serebral) yang terjadi karena berkurangnya aliran darah dan oksigen ke otak. Berkurangnya aliran darah dan oksigen ini bisa dikarenakan adanya

sumbatan, penyempitan atau pecahnya pembuluh darah (Pinzon dan Asanti, 2010).

Di Indonesia, insidensi dan prevalensi *stroke* belum diketahui secara pasti. Diperkirakan 500.000 penduduk terkena *stroke* setiap tahunnya, sekitar 2,5% atau 12.500 orang meninggal dan sisanya cacat ringan. Hampir setiap hari, atau minimal rata-rata tiga hari sekali ada seorang penduduk Indonesia baik tua maupun muda meninggal dunia karena serangan *stroke* (Depkes, 2009).

Data Mining

Data mining sering juga disebut *knowledge discovery in database* (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari *data mining* bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan. *Data mining* juga meliputi langkah-langkah menentukan variabel atau fitur yang penting untuk dipakai dalam klasifikasi dan regresi. . *Data mining* memegang peran penting dalam bidang industri, keuangan, cuaca, ilmu pengetahuan dan teknologi. *Data mining* berkenaan dengan pengolahan data dalam skala besar (Santosa, 2007).

Klasifikasi

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu pembangunan model sebagai prototipe untuk disimpan sebagai memori dan penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya (Prasetyo, 2014).

Data *Training* dan Data *Testing*

Menurut Prasetyo (2014), data untuk pengujian klasifikasi dibagi menjadi data *training* dan data *testing.* Data atau vektor yang sudah diketahui sebelumnya untuk label kelas dan digunakan untuk membangun model *classifier* disebut dengan data *training.* Data atau vektor yang belum diketahui (dianggap belum diketahui) label kelasnya menggunakan model *classifier* yang sudah dibangun disebut data *testing.*

***Naive* Bayes**

Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik sederhana yang berdasar pada teorema Bayes dengan asumsi independensi yang kuat (Prasetyo, 2014). *Naive* Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* dengan data yang besar (Kusrini dan Luthfi, 2009). Definisi lain mengatakan *naive* Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Bustami, 2013). Persamaan dari Teorema Bayes adalah :

$$P(C \mid F) = \frac{P(C) \times P(F \mid C)}{P(F)} \qquad (1)$$

di mana :

F	:	Data dengan kelas yang belum diketahui
C	:	Hipotesis data merupakan suatu kelas spesifik
$P(C \mid F)$	:	Probabilitas hipotesis C dengan syarat F (probabilitas posterior)
$P(C)$	:	Probabilitas hipotesis C (probabilitas prior)
$P(F \mid C)$	:	Probabilitas hipotesis F dengan syarat C
$P(F)$	:	Probabilitas hipotesis F

Untuk menjelaskan metode *naive* Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, metode *naive* Bayes di atas disesuaikan sebagai berikut :

$$P(C|F_1 \dots F_n) = \frac{P(C).P(F_1 \dots F_n \mid C)}{P(F_1 \dots F_n)} \qquad (2)$$

Di mana variabel C mempresentasikan kelas, sementara variabel $F_1 \dots F_n$ mempresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel dengan karakteristik tertentu dalam kelas C (*posterior*) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut *prior*), dikalikan dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (disebut juga *likelihood*), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara umum (disebut juga *evidence*). Karena itu, rumus dapat pula ditulis secara sederhana sebagai berikut :

$$\text{Posterior} = \frac{prior \times likelihood}{evidence} \qquad (3)$$

Nilai *evidence* selalu tetap untuk setiap kelas pada satu sampel (Prasetyo, 2014).

Penjabaran lebih lanjut rumus Bayes tersebut dilakukan dengan menjabarkan $(C \mid F_1, \dots, F_2)$ menggunakan aturan perkalian sebagai berikut :

$$P(C|F_1 \dots F_n) = P(C)P(F_1,\dots,F_n|C)$$
$$= P(C)P(F_1|C)P(F_2,\dots,F_n|C,F_1)$$
$$= P(C)P(F_1|C)P(F_2|C,F_1)P(F_3,\dots,F_n \mid C,F_1,F_2)$$
$$= P(C)P(F_1|C)P(F_2|C,F_1)P(F_3|C,F_1,F_2)...P(F_n \mid C,F_1,F_2,F_3,...,F_{n-1}) \quad (4)$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor-faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dilakukan analisis satu persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Disinilah digunakan asumsi independensi yang sangat tinggi (*naive*), bahwa masing-masing petunjuk $(F_1,F_2,...,F_n)$ saling bebas (*independent*) satu sama lain. Dengan asumsi tersebut maka berlaku suatu kesamaan sebagai berikut:

$$P(F_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i)$$

untuk $i \neq j$, (5)

Sehingga,

$$P(F_i|C,F_j) = P(F_i|C) \quad (6)$$

Persamaan di atas dapat disimpulkan bahwa asumsi independensi *naive* tersebut membuat syarat peluang menjadi sederhana, sehingga perhitungan menjadi mungkin untuk dilakukan. Selanjutnya, penjabaran $P(C \mid F_1,\dots,F_n)$ dapat disederhanakan menjadi :

$$P(C|F_1,\dots,F_n) = P(C)P(F_1|C)P(F_2|C)P(F_3|C)\dots P(F_n|C)$$
$$= P(C)\prod_{i=1}^{n} P(F_i \mid C) \quad (7)$$

***Decision Tree* Algoritma (J48)**

Algoritma *decision tree* merupakan algoritma yang umum digunakan untuk pengambilan keputusan. *Decision tree* akan mencari solusi permasalahan dengan menjadikan kriteria sebagai *node* yang saling berhubungan membentuk seperti struktur pohon. *Decision tree* adalah model prediksi terhadap suatu keputusan menggunakan struktur hirarki atau pohon. Setiap pohon memiliki cabang, cabang mewakili setiap atribut yang harus dipenuhi untuk menuju cabang selanjutnya hingga berakhir di daun (tidak ada cabang lagi). *Decision tree* algoritma (J48) merupakan implementasi dari algoritma (C4.5) yang memproduksi *decision tree*. Ini merupakan standar algoritma yang digunakan dalam *machine learning*. *Decision tree* merupakan salah satu algoritma klasifikasi dalam *data mining*.

Secara umum Algoritma (J48) untuk membangun *decision tree* adalah sebagai berikut (Kusrini dan Luthfi, 2009) :

1. Pemilihan Atribut Akar

   Untuk memilih atribut akar, digunakan nilai *gain ratio* dari atribut-atribut yang ada. Berikut adalah cara untuk menghitung nilai *gain ratio* :

$$Gain\ Ratio(S,A) = \frac{Gain(S,A)}{Split\ Info(S,A)} \quad (8)$$

dengan

S : Himpunan kasus
A : Atribut
$Gain(S,A)$: *Information gain* pada atribut A
$Split\ Info(S,A)$: *Split information* pada atribut A

Sebelum mendapatkan nilai *gain ratio*, dicari terlebih dahulu nilai *gain* dan *split information* nya. Berikut adalah cara untuk menghitung nilai *gain* :

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{m} \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (9)$$

dengan:

S : Himpunan kasus
S_i : Himpunan kasus pada partisi ke-i
A : Atribut
m : Jumlah partisi
$|S_i|$: Jumlah kasus pada partisi ke-i
$|S|$: Jumlah kasus dalam S

Untuk mendapatkan nilai *gain*, dicari terlebih dahulu nilai *entropy*. *Entropy* adalah informasi mengenai proporsi pembagian kelas, nilai *entropy* berkisar mulai dari 0 sampai dengan 1, jika nilai *entropy* = 0, maka menandakan jumlah sampel hanya berada di salah satu kelas, sedangkan jika nilai *entropy* = 1, maka menandakan jumlah sampel berada di masing-masing kelas dengan jumlah yang sama. Adapun rumus dasar dari perhitungan *entropy* adalah sebagai berikut :

$$Entropy(S) = -\sum_{i=1}^{m} p_i log_2 p_i \quad (10)$$

dengan

S : Himpunan Kasus
m : Jumlah Partisi
p_i : Proporsi dari S_i terhadap S

Kemudian untuk menghitung *gain ratio* kita perlu menghitung *split information*. *Split information* dihitung dengan persamaan sebagai berikut :

$$Split\ Info(S,A) = -\sum_{i=1}^{m} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \quad (11)$$

dengan

S : Himpunan kasus
A : Atribut
m : Jumlah partisi
S_i : Himpunan kasus pada partisi ke-i

$|S_i|$: Jumlah kasus pada partisi ke-i

$|S|$: Jumlah kasus dalam S

2. Penentuan Cabang untuk Masing-masing Nilai Untuk penentuan cabang didasarkan pada nilai *gain ratio* tertinggi dari atribut-atribut yang ada.
3. Kelas dibagi dalam cabang dan apabila cabang mempunyai dua kelas maka dipilih kelas yang terbanyak.
4. Proses diulang untuk masing-masing cabang sampai semua kelas pada cabang memiliki kelasnya masing masing.

Pengukuran Kinerja Klasifikasi

Menurut Prasetyo (2014), sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua set data dengan benar. Akan tetapi, tidak dipungkiri bahwa kinerja suatu sistem tidak bisa bekerja 100% benar. Oleh karena itu, sebuah sistem klasifikasi juga harus diukur kinerjanya. Untuk menghitung akurasi digunakan persamaan sebagai berikut:

$$\text{akurasi} = \frac{\text{jumlah data yang diprediksi secara benar}}{\text{jumlah prediksi yang dilakukan}} \times 100\% \quad (12)$$

WEKA (*Waikato Environment for Knowledge Analysis*)

Waikato Environment for Knowledge Analysis (WEKA) merupakan perangkat lunak pembelajaran mesin yang populer yang ditulis dalam bahasa pemrograman java. WEKA dikembangkan oleh Universitas Waikato di Selandia Baru. WEKA berisikan kumpulan algoritma beserta visualisasinya untuk analisis data dan pemodelan prediktif. Algoritma-algoritma pembelajaran mesin pada WEKA dapat dimanfaatkan untuk pemecahan masalah di bidang *data mining*. WEKA memiliki implementasi semua teknik pembelajaran untuk klasifikasi dan regresi, yaitu *decision tree*, *rules set*, pengklasifikasian teorema Bayes, *Support Vector Machines* (SVM), logistik dan linier, *multi layers perceptrons* dan metode *nearest neighbour* (Sartika, 2017).

Pengertian *Stroke*

Stroke didefinisikan sebagai gangguan fungsi sistem saraf yang terjadi mendadak dan disebabkan oleh gangguan peredaran darah otak. Gangguan peredaran darah otak dapat berupa tersumbatnya pembuluh darah otak atau pecahnya pembuluh darah di otak. Adapun gejala yang ditimbulkan oleh penyakit *stroke* adalah kelumpuhan anggota gerak, wajah perot, gangguan bicara, pusing berputar, nyeri kepala dan penurunan kesadaran (Pinzon dan Asanti, 2010).

Menurut Pudiastuti (2011) *stroke* terbagi menjadi 2 kategori yaitu *stroke* hemoragik (pendarahan) dan *stroke* non hemoragik (non pendarahan).

1. *Stroke* Hemoragik adalah *stroke* karena pecahnya pembuluh darah sehingga menghambat aliran darah yang normal dan darah merembes ke dalam daerah otak dan merusaknya. Hampir 70% kasus *stroke* hemoragik diderita oleh penderita hipertensi.
2. *Stroke* non hemoragik terjadi karena tersumbatnya pembuluh darah yang menyebabkan aliran darah ke otak sebagian atau keseluruhan terhenti. Hal ini disebabkan oleh aterosklerosis yaitu penumpukan kolesterol pada dinding pembuluh darah atau bekuan darah yang telah menyumbat suatu pembuluh darah ke otak.

Stroke merupakan suatu penyakit yang disebabkan oleh banyak faktor risiko atau biasa disebut multikausal (Wahjoepramono, 2005). Seseorang menderita *stroke* karena memiliki faktor risiko *stroke*. Faktor risiko dapat dibagi menjadi dua, yaitu faktor risiko yang tidak dapat diubah (umur, jenis kelamin, ras, riwayat keluarga dan riwayat *stroke* sebelumnya) dan faktor risiko yang dapat diubah (hipertensi, diabetes melitus, dislipidemia, merokok dan obesitas).

Hasil dan Pembahasan

1. Statistika Deskriptif

Tahap awal yang dilakukan dalam penelitian ini adalah analisis deskriptif, bertujuan untuk menggambarkan karakteristik data pasien rawat inap dari RSUD Abdul Wahab Sjahranie Samarinda bulan November dan Desember 2017. Karakteristik yang digambarkan pada analisis deskriptif adalah umur, jenis kelamin, tekanan darah, diabetes melitus, dislipidemia, kadar asam urat dan penyakit jantung.

2. Klasifikasi *Naive* Bayes

Dalam proses menghitung klasifikasi *naive* Bayes, terdapat tiga alur yaitu membaca data *training*, menghitung nilai probabilitas setiap atribut pada setiap kelasnya dan menentukan probabilitas akhir. Adapun data yang digunakan untuk proses perhitungan klasifikasi adalah dengan *data training* 90% sebanyak 140 sampel, sedangkan untuk *data testing* 10% sebanyak 16 sampel. Probabilitas akhir yang didapat digunakan untuk menentukan termasuk dalam kelas *stroke* hemoragik atau *stroke* non hemoragik untuk data yang ingin diuji. Misalkan, sebagai contoh kasus akan memprediksi kelas pada seseorang dengan atribut pasien yang memiliki umur kategori lansia, berjenis kelamin perempuan, memiliki tekanan darah tinggi, tidak memiliki diabetes melitus, memiliki dislipidemia normal, memiliki kadar asam urat tinggi dan memiliki penyakit jantung.

Selanjutnya akan ditentukan apakah pasien tersebut termasuk kelas *stroke* hemoragik atau *stroke* non hemoragik.

Untuk menentukan data yang akan dianalisis dengan metode *naive* Bayes maka tahap pertama yang dilakukan adalah membaca *data training*. Tahap kedua adalah menentukan nilai probabilitas setiap atribut pada setiap kelasnya (*likelihood*) dengan data yang digunakan adalah data kualitatif. Pada penelitian ini, atribut yang digunakan adalah umur (F_1), jenis kelamin (F_2), tekanan darah (F_3), diabetes melitus (F_4), dislipidemia (F_5), kadar asam urat (F_6), penyakit jantung (F_7) dan status *stroke* (C).

Adapun nilai probabilitas atribut umur pada setiap kelasnya adalah sebagai berikut :

Umur (F_1)

Pada perhitungan nilai probabilitas atribut umur pada kelas hemoragik terdiri dari 11 pasien rawat inap yang memiliki umur dengan kategori dewasa dan 60 pasien rawat inap yang memiliki umur dengan kategori lansia, sedangkan pada kelas non hemoragik terdiri dari 50 pasien rawat inap yang memiliki umur dengan kategori dewasa dan 19 pasien rawat inap yang memiliki umur dengan kategori lansia. Adapun nilai probabilitas atribut umur pada setiap kelasnya yaitu sebagai berikut:

Probabilitas atribut umur kategori dewasa untuk kelas hemoragik, dihitung sebagai berikut,

$$P(\text{Umur = Dewasa}|\text{Hemoragik}) = \frac{11}{11+60} = 0{,}1549$$

Jadi, peluang pasien dengan umur kategori dewasa terkena penyakit *stroke* hemoragik sebesar 0,1549.

Probabilitas atribut umur kategori lansia untuk kelas hemoragik, dihitung sebagai berikut,

$$P(\text{Umur = Lansia}|\text{Hemoragik}) = \frac{60}{60+11} = 0{,}8451$$

Jadi, peluang pasien dengan umur kategori lansia terkena penyakit *stroke* hemoragik sebesar 0,8451.

Probabilitas atribut umur kategori dewasa untuk kelas non hemoragik, dihitung sebagai berikut,

$$P(\text{Umur = Dewasa}|\text{Non Hemoragik}) = \frac{50}{50+19} = 0{,}7246$$

Jadi, peluang pasien dengan umur kategori dewasa terkena penyakit *stroke* non hemoragik sebesar 0,7246.

Probabilitas atribut umur kategori lansia untuk kelas non hemoragik, dihitung sebagai berikut,

$$P(\text{Umur = Lansia}|\text{Non Hemoragik}) = \frac{19}{19+50} = 0{,}2754$$

Jadi, peluang pasien dengan umur kategori lansia terkena penyakit *stroke* non hemoragik sebesar 0,2754. Adapun nilai probabilitas atribut umur pada setiap kelasnya dapat dilihat pada Tabel 1.

Tabel 1 Probabilitas Umur pada Setiap Kelasnya

Umur	Status *Stroke* (Kelas)		Probabilitas Status *Stroke* (Kelas)	
	H	NH	H	NH
Dewasa	11	50	0,1549	0,7246
Lansia	60	19	0,8451	0,2754

Keterangan:
H : Hemoragik
NH : Non Hemoragik

Dengan cara yang sama diperoleh nilai probabilitas atribut jenis kelamin (F_2), tekanan darah (F_3), diabetes melitus (F_4), dislipidemia (F_5), kadar asam urat (F_6), penyakit jantung (F_7) dan status *stroke* (C). Adapun nilai probabilitas seluruh atribut pada setiap kelasnya dapat dilihat pada Tabel 2.

Tabel 2. Probabilitas Setiap Atribut pada Setiap Kelasnya

Atribut	Kategori	Probabilitas Status *Stroke* (Kelas)	
		H	NH
Jenis Kelamin	Laki-laki	0,5070	0,6957
	Perempuan	0,4930	0,3043
Tekanan Darah	Normal	0,0423	0,5362
	Tinggi	0,9577	0,4638
Diabetes melitus	Ada	0,2535	0,9275
	Tidak Ada	0,7465	0,0725
Dislipi-Demia	Normal	0,6901	0,1449
	Tinggi	0,3099	0,8551
Kadar Asam Urat	Normal	0,4930	0,5942
	Tinggi	0,5070	0,4058
Penyakit Jantung	Ada	0,4648	0,6667
	Tidak Ada	0,5352	0,3333
Status *Stroke* (C)		0,5071	0,4929

Keterangan:
H : Hemoragik
NH : Non Hemoragik

Setelah mengetahui probabilitas setiap atribut pada setiap kelasnya maka selanjutnya ke tahap tiga adalah menentukan probabilitas akhir (peluang posterior). Perhitungan probabilitas akhir setiap kelas menggunakan Persamaan (7) sebagai berikut:

$\prod_{i=1}^{7} P(\text{Hemoragik} \mid F_1, F_2, F_3, F_4, F_5, F_6, F_7) = P(\text{Hemoragik}) \times P(F_1 \mid \text{Hemoragik}) \times P(F_2 \mid \text{Hemoragik}) \times P(F_3 \mid \text{Hemoragik}) \times P(F_4 \mid \text{Hemoragik}) \times P(F_5 \mid \text{Hemoragik}) \times P(F_6 \mid \text{Hemoragik}) \times P(F_7 \mid \text{Hemoragik})$

$= P(\text{Hemoragik}) \times P(\text{Umur = Lansia} \mid \text{Hemoragik}) \times P(\text{Jenis Kelamin = Perempuan} \mid \text{Hemoragik}) \times P(\text{Tekanan Darah = Tinggi} \mid \text{Hemoragik}) \times P(\text{Diabetes Melitus =Tidak Ada} \mid \text{Hemoragik}) \times P(\text{Dislipidemia = Normal} \mid \text{Hemoragik}) \times P(\text{Kadar AsamUrat =Tinggi} \mid \text{Hemoragik}) \times P(\text{Penyakit Jantung = Ada} \mid \text{Hemoragik})$

$= 0{,}5071 \times 0{,}0484$

$= 0{,}0245$

$\prod_{i=1}^{7} P(\text{Non Hemoragik} \mid F_1, F_2, F_3, F_4, F_5, F_6, F_7) = P(\text{Non Hemoragik}) \times P(F_1 \mid \text{Non Hemoragik}) \times P(F_2 \mid \text{Non Hemoragik}) \times P(F_3 \mid \text{Non Hemoragik}) \times P(F_4 \mid \text{Non Hemoragik}) \times P(F_5 \mid \text{Non Hemoragik}) \times P(F_6 \mid \text{Non Hemoragik}) \times P(F_7 \mid \text{Non Hemoragik})$

$= P(\text{Non Hemoragik}) \times P(\text{Umur = Lansia} \mid \text{Non Hemoragik}) \times P(\text{Jenis Kelamin = Perempuan} \mid \text{Non Hemoragik}) \times P(\text{Tekanan Darah = Tinggi} \mid \text{Non Hemoragik}) \times P(\text{Diabetes Melitus = Tidak Ada} \mid \text{Non Hemoragik}) \times P(\text{Dislipidemia = Normal} \mid \text{Non Hemoragik}) \times P(\text{Kadar AsaM Urat =Tinggi} \mid \text{Non Hemoragik}) \times P(\text{Penyakit Jantung = Ada} \mid \text{Non Hemoragik})$

$= 0{,}4929 \times 0{,}0001$

$= 0{,}00005$

Berdasarkan perhitungan di atas, dapat diketahui bahwa kelas yang memiliki nilai probabilitas terbesar adalah kelas hemoragik, sehingga dapat disimpulkan kasus dengan nilai-nilai atribut umur pasien rawat inap kategori lansia, berjenis kelamin perempuan, tekanan darah tinggi, tidak memiliki diabetes melitus, dislipidemia normal, kadar asam urat tinggi dan memiliki penyakit jantung diprediksi masuk dalam kelas hemoragik artinya pasien tersebut masuk kelas terkena penyakit *stroke* hemoragik (pendarahan).

3. Klasifikasi *Decision Tree* Algoritma (J48)

Dalam proses pembentukan pohon klasifikasi terdapat tiga alur yaitu penentuan *node* akar, menentukan cabang untuk masing-masing atribut dan proses diulang untuk masing-masing cabang sampai semua kelas pada cabang memiliki kelasnya masing-masing.

Tahap pertama dalam pembentukan pohon klasifikasi adalah pemilihan *node* akar. Perhitungan *node* akar menggunakan Persamaan (8) untuk menentukan *gain ratio*, menghitung nilai *gain* menggunakan Persamaan (9), menghitung nilai *entropy* menggunakan Persamaan (10) dan menghitung nilai *split information* menggunakan Persamaan (11). Atribut yang digunakan untuk menentukan *node* akar adalah umur (F_1), jenis kelamin (F_2), tekanan darah (F_3), diabetes melitus (F_4), dislipidemia (F_5), kadar asam urat (F_6) dan penyakit jantung (F_7). Adapun hasil perhitungan *gain ratio* berdasarkan Persamaan (8), *gain* berdasarkan Persamaan (9), *entropy* berdasarkan Persamaan (10) dan *split information* berdasarkan Persamaan (11) pada atribut umur adalah sebagai berikut :

***Entropy* Total**

$$Entropy\ (S) = -\sum_{i=1}^{2} p_i \log_2 p_i$$

$$Entropy\ (S) = \left(-\frac{71}{140} \times \log_2 \frac{71}{140}\right) + \left(-\frac{69}{140} \times \log_2 \frac{69}{140}\right)$$
$$= \left(-0{,}5071 \times (-0{,}9795)\right) + \left(-0{,}4929 \times (-1{,}0208)\right)$$
$$= 0{,}9999$$

***Entropy* Umur**

$$Entropy\ (\text{Dewasa}) = \left(-\frac{11}{61} \times \log_2 \frac{11}{61}\right) + \left(-\frac{50}{61} \times \log_2 \frac{50}{61}\right)$$
$$= \left(-0{,}1803 \times (-2{,}4713)\right) + \left(-0{,}8197 \times (-0{,}2869)\right)$$
$$= 0{,}6808$$

$$Entropy\ (\text{Lansia}) = \left(-\frac{60}{79} \times \log_2 \frac{60}{79}\right) + \left(-\frac{19}{79} \times \log_2 \frac{19}{79}\right)$$
$$= \left(-0{,}7595 \times (-0{,}3969)\right) + \left(-0{,}2405 \times (-2{,}0559)\right)$$
$$= 0{,}7959$$

***Gain* Umur**

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{2} \frac{|S_i|}{|S|} \times Entropy(S_i)$$

$$Gain\ (\text{Umur}) = 0{,}9999 - \left(\left(\frac{61}{140} \times 0{,}6808\right) + \left(\frac{79}{140} \times 0{,}7959\right)\right)$$
$$= 0{,}2541$$

***Split Information* Umur**

$$Split\ Info(S, A) = -\sum_{i=1}^{2} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$Split\ Info(\text{Umur}) = \left(\left(-\frac{61}{140} \times \log_2 \frac{61}{140}\right) + \left(-\frac{79}{140} \times \log_2 \frac{79}{140}\right)\right)$$
$$= 0{,}9880$$

***Gain Ratio* Umur**

$$Gain\ Ratio(S, A) = \frac{Gain(S, A)}{Split\ Info(S, A)}$$

$$Gain\ Ratio\ (\text{Umur}) = \frac{Gain\ (\text{Umur})}{Split\ Info\ (\text{Umur})}$$
$$= \frac{0,2541}{0,9880}$$
$$= 0,2572$$

Adapun hasil perhitungan *gain ratio* berdasarkan Persamaan (8), *gain* berdasarkan Persamaan (9), *entropy* berdasarkan Persamaan (10) dan *split information* berdasarkan Persamaan (11) seluruh atribut pada setiap kelasnya dapat dilihat pada Tabel 3.

Tabel 3. Hasil Perhitungan *Entropy*, *Gain*, *Split Information* dan *Gain Ratio* untuk *Node* Akar untuk *Node* 1

Atribut	Kategori	*E*	*G*	*SI*	*GR*
Total		**0,99985**			
Umur	Dewasa	0,6808	0,2541	0,9880	0,2572
	Lansia	0,7959			
Jenis Kelamin	Laki-laki	0,9852	0,0269	0,9710	0,0277
	Perempuan	0,9544			
Tekanan Darah	Normal	0,3843	0,2441	0,8631	0,2828
	Tinggi	0,9044			
Diabetes Melitus	**Ada**	**0,7593**	**0,3796**	**0,9787**	**0,3879**
	Tidak Ada	**0,4237**			
Dislipi-demia	Normal	0,6565	0,2350	0,9821	0,2393
	Tinggi	0,8438			
Kadar Asam Urat	Normal	0,9955	0,0075	0,9947	0,0075
	Tinggi	0,9887			
Penyakit Jantung	Ada	0,9804	0,0301	0,9880	0,0305
	Tidak Ada	0,9559			

Keterangan:
E : *Entropy*
G : *Gain*
SI : *Split Information*
GR : *Gain Ratio*

Berdasarkan Tabel 3 menunjukkan bahwa *gain ratio* tertinggi ada di atribut diabetes melitus sehingga atribut diabetes melitus dijadikan sebagai *node* akar (*node* 1). Maka cabang untuk *node* akar ada dua, yaitu {Ada} dan {Tidak Ada}.

Selanjutnya untuk cabang *node* pada *node* 2, *node* 3, *node* 4 dan seterusnya, *gain*, *entropy*, *split information* dan *gain ratio* dihitung terlebih dahulu seperti pada langkah awal mencari *node* akar namun data yang disesuaikan dengan *node* yang akan dihitung *gain ratio* nya. Perhitungan untuk menentukan *node* 2 dan seterusnya menggunakan *gain ratio* berdasarkan Persamaan (8), *gain* berdasarkan Persamaan (9), *entropy* berdasarkan Persamaan (10) dan *split information* berdasarkan Persamaan (11). Terus berlanjut sampai ke *node* di mana semua *node* tidak ada penurunan lagi.

Uji Akurasi *Naive* Bayes dan *Decision Tree* Algoritma (J48)

Dalam melakukan prediksi (klasifikasi) diharapkan dapat melakukan klasifikasi pada semua objek dengan benar. Pada mode penelitian ini data set sebanyak 156 data akan dibagi menjadi *data training* dan *data testing* sesuai persentase yang ditentukan. Persentase yang digunakan dalam mode pengujian ini adalah dengan data *training* sebesar 90% dari data set. Untuk mendapatkan tingkat akurasi (ketepatan) menggunakan Persamaan (12). Adapun hasil akurasi dapat dilihat pada Tabel 4.

Tabel 4. Hasil Akurasi Klasifikasi

Metode	Akurasi(%)
Decision Tree J48	87,5
Naive Bayes	81,25

Kesimpulan

Berdasarkan hasil analisis dan pembahasan, dapat disimpulkan bahwa hasil ketepatan klasifikasi penyakit *stroke* pada data pasien di RSUD Abdul Wahab Sjahranie bulan November dan Desember 2017 dengan metode *naive* Bayes adalah 81,25% dan metode *decision tree* algoritma (J48) diperoleh tingkat akurasi sebesar 87,5%. Hal ini menunjukkan bahwa pada penelitian ini, metode *decision tree* algoritma (J48) memberikan ketepatan prediksi klasifikasi yang lebih baik.

Daftar Pustaka

Bustami. (2013). Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi. *TECHSI : Jurnal Penelitian Teknik Informatika*, 3(2), 129-132.

Departemen Kesehatan RI. (2009). *Pedoman Pelayanan Antenatal di Tingkat Pelayanan Dasar*. Jakarta : Depkes RI.

Induniasih dan Ratna, W. (2017). *Promosi Kesehatan*. Yogyakarta : Pustaka Baru Press.

Kemenkes RI. (2012). *Buletin Jendela Data dan Informasi Kesehatan Penyakit Tidak Menular*. Jakarta.

Kusrini dan Luthfi, E. T. (2009). *Algoritma Data Mining*. Yogyakarta : Andi Offset.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey : John Wiley & Sons.

Pinzon, R. dan Asanti, L. (2010). *Awas Stroke : Pengertian, Gejala, Tindakan, Perawatan, & Pencegahan*. Yogyakarta : Andi Offset.

Prasetyo, E. (2014). *Data Mining: Konsep Dan Aplikasi Menggunakan Matlab*. Yogyakarta : Andi Offset

Pudiastuti, Ratna D. (2011). *Penyakit Pemicu Stroke*. Yogyakarta: Nuha Medika.

Sartika, D. (2017). Perbandingan Algoritma Klasifikasi *Naive* Bayes, *Nearest Neighbour*, dan *Decision Tree* pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian. *Jatisi*, 2(1), 151-161.

Wahjoepramono. (2005). *Stroke Tata Laksana Fase Akut.* Jakarta : Fakultas Kedokteran Universitas Pelita Harapan, RS Siloam Gleneagles.