The Effect of Using Dummy Variable on Classification of Womb Disease with C4.5 Method

Moch Shofieyuddin
Informatika, Fakultas MIPA
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36 A Surakarta
moch.shofieyuddin@student.uns.ac.id

Ristu Saptono
Informatika, Fakultas MIPA
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36 A Surakarta
ristu.saptono@staff.uns.ac.id

Afrizal Doewes
Informatika, Fakultas MIPA
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36 A Surakarta
afrizal.doewes@staff.uns.ac.id

ABSTRACT

The use of dummy variables is recommended because the symptoms of the womb disease compounds that have the possible values that appear more than two (non-binary), there is a possibility that not all types of occurrence related to the disease symptoms as other content that needs to be done solving the symptoms so that the value to binary and symptoms become more specific. By applying the dummy variable, is expected to improve the accuracy of the probabilistic approach Naïve Bayes classifier, because the assumption of independency between the symptoms of the disease are met. Besides Naïve Bayes classifier, Decission Tree is also commonly used in classification, one of Decission Tree method is C4.5. This study discusses the effect of the use of dummy variables in the womb disease classification using C4.5. From the results of this study concluded that the use of dummy variables to produce an average value accuracy, precission, recall, and F-measure which remained stable at 87.2% in testing k-fold cross validation with value of k (5, 10, 15, 20, and 25). However, the use of dummy variables reduces the average value of accuracy, precission, recall, and F-measure sequentially from 89.6%, 89.74%, 89.7%, and 89.6% to 87.2%, 87.2%, 87.2% and 87.2%. Besides, the use of dummy variables to specify the attributes of disease symptoms used in the classification of disease womb.

Keywords: *Dummy Variable, C4.5 method, Womb Disease.*

1. PENDAHULUAN

Dalam *statistic* dan *econometric*, khususnya dalam *regression analysis*, *dummy variable* adalah variabel buatan dibuat untuk mewakili sebuah atribut dengan dua atau lebih kategori yang berbeda. Dibutuhkan nilai 0 atau 1 untuk menunjukkan ketiadaan atau kehadiran beberapa efek kategoris yang dapat diharapkan untuk mengubah hasil output. *Dummy variable* digunakan sebagai perangkat untuk mengurutkan data ke dalam kategori yang saling eksklusif. Misalnya, dalam *economic time series analysis*, *dummy variable* digunakan untuk menunjukkan terjadi atau tidak terjadinya peristiwa [1].

Dalam penelitian [2] tentang klasifikasi penyakit kandungan dengan menggunakan data dari penelitian yang dilakukan oleh [3] disarankan penggunaan *dummy variable* untuk pendekatan *Naïve Bayes Classifier*. Penggunaan *dummy variable* disarankan karena pada gejala penyakit kandungan yang memiliki kemungkinan nilai yang muncul lebih dari dua (*non-binary*), ada kemungkinan tidak semua jenis kemunculannya terkait pada gejala penyakit kandungan lainnya, sehingga perlu dilakukan pemecahan gejala penyakit kandungan agar nilainya menjadi *binary* dan gejala menjadi lebih spesifik. Dengan mengaplikasikan *dummy variable* diharapkan dapat meningkatkan akurasi pada pendekatan probabilistik *Naïve Bayes Classifier*, karena asumsi indepedensi antar gejala penyakit terpenuhi.

Selain *Naïve Bayes Classifier, Decission Tree* juga sering digunakan dalam klasifikasi, salah satu metode *Decission Tree* yaitu metode C4.5. Pada penelitian yang dilakukan [4] tentang teknik *data mining* dalam diagnosis dan prognosis pada penyakit kanker, menyimpulkan bahwa metode C4.5 lebih baik dibandingkan teknik *data mining* lainnya, salah satunya adalah *Naïve Bayes Classifier*.

Berdasarkan penelitian diatas, dengan menggunakan data set [3] akan dilakukan penelitian menggunakan pendekatan yang berbeda. Pada penelitian ini akan dibandingkan pengaruh penggunaan *Dummy Variable* terhadap klasifikasi penyakit kandungan dengan menggunakan metode C4.5.

2. DASAR TEORI

2.1. Decission Tree

Decision tree adalah struktur *flowchart* yang menyerupai *tree* (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas [5]. Alur pada *decision tree* di telusuri dari simpul akar ke simpul daun yang memegang prediksi kelas untuk contoh tersebut. Hasil penelusuran dari simpul akar hingga simpul daun akan membentuk *rule-rule* yang digunakan untuk klasifikasi.

2.2. C4.5

Algoritma C4.5 dan pohon keputusan merupakan dua model yang tak terpisahkan. Algoritma C4.5 merupakan salah satu algoritma klasifikasi yang kuat dan cukup banyak digunakan atau di implementasikan untuk pengklasifikasian dalam berbagai hal.

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 (*Iterative Dichotomiser Tree*).

Serangkaian perbaikan yang dilakukan pada algoritma ID3 mencapai puncaknya dengan menghasilkan sebuah sistem praktis dan berpengaruh untuk pembentukan pohon keputusan. Perbaikan tersebut meliputi metode untuk menangani data numerik *attributes*, *missing values*, *noisy data*, dan aturan yang menghasilkan aturan dari *tree* [6].

Saat menyusun sebuah pohon keputusan pertama yang harus dilakukan adalah menentukan atribut mana yang akan menjadi simpul akar dan atribut mana yang akan menjadi simpul selanjutnya. Dengan penambahan penghitungan *gain ratio* yaitu informasi paling potensial dari seluruh entropi. *Gain Ratio* merupakan modifikasi dari *information gain* untuk mengurangi bias atribut yang memiliki banyak cabang. *Gain ratio* meningkatkan keakurasian pengambilan *information gain* dari algoritma ID3. Pada penghitungan *gain* ratio, dilakukan penghitungan *split info* yaitu informasi potensial dari hasil tes entrophy suatu atribut yang bertujuan untuk memisahkan kelas target. Secara sederhana algoritma C4.5 dapat dilakukan dengan langkah-langkah pada Algoritma 1.

Algoritma 1. Algoritma C4.5

- *Input*: data *training*
- *Output*: pohon keputusan C4.5

- Langkah-langkah:

1. Menghitung *information gain* dari setiap atribut dengan menggunakan:

$$Gain(S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{Si}{S} \times Entropy(Si) \quad (2.1)$$

dimana,

$$Entropy(S) = \sum_{i=1}^{n} - p_i\ log_2\ p_i \quad (2.2)$$

Keterangan:
S = himpunan kasus
n = banyaknya partisi S
pi = probabilitas yang didapat dari kelas dibagi total kasus
A = semua nilai yang mungkin dari atribut A
$S\ i$ = subset dari y dimana A mempunyai nilai i

2. Menghitung *gain ratio* dengan menggunakan persamaan:

$$Gain\ Ratio\ (S, A) = \frac{Gain\ (S,A)}{SplitInfo\ (S,A)} \quad (2.3)$$

dimana,

$$SplitInfo(S, A) = -\sum_{i=1}^{n} \frac{Si}{S} Log_2 \frac{Si}{S} \quad (2.4)$$

3. Memilih atribut yang memiliki nilai *gain ratio* terbesar sebagai *node* awal atau *root*

4. Mengulangi perhitungan 1 dan 2 untuk atribut yang belum terpilih dengan mengikutsertakan masing-masing kelas atribut *root*

5. Mengulangi seluruh langkah hingga seluruh kelas masuk ke dalam tree.
6. Memangkas 1 cabang dengan probabilitas kemunculan kelas terendah, jika akurasi setelah dipangkas lebih tinggi maka cabang tersebut dipangkas, jika tidak cabang tersebut dibiarkan.

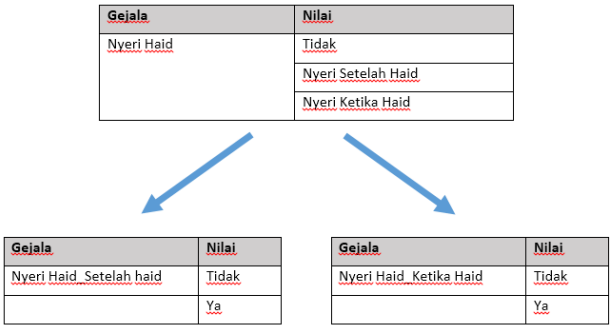
7. Mengulangi langkah 6 ke seluruh cabang hingga terbentuk *tree* yang lebih sederhana

2.3. *Dummy Variable*

Dalam *statistic* dan *econometric*, khususnya dalam *regression analysis*, *dummy variable* adalah variabel buatan dibuat untuk mewakili sebuah atribut dengan dua atau lebih kategori yang berbeda. Dibutuhkan nilai 0 atau 1 untuk menunjukkan ketiadaan atau kehadiran beberapa efek kategoris yang dapat diharapkan untuk mengubah hasil output. *Dummy variable* digunakan sebagai perangkat untuk mengurutkan data ke dalam kategori yang saling eksklusif. Misalnya, economic time series analysis, *dummy variable* digunakan untuk menunjukkan terjadi atau tidak terjadinya peristiwa [1].

Dummy variable merupakan pengkodean ulang dari *categorical variables* yang mempunyai lebih dari dua kategori yang diubah menjadi beberapa *binary variable.* Contoh : Status Pernikahan, jika data asli dilabeli dengan 1 = Menikah, 2 = Belum Menikah, 3 = Cerai/Janda/ Duda/Berpisah, dapat diubah menjadi dua variable sebagai berikut : *var_1* : 1 = Belum Menikah, 0 = Lain, *var_2* : 1 = Cerai/Janda/ Duda/Berpisah.

Untuk kasus diatas jika ada seorang yang sudah menikah, maka kedua *var_1* dan *var_2* akan memiliki nilai 0. Umumnya, *categorical variable* dengan kategori (*k*) akan dikodekan mennjadi (*k -1*) untuk *dummy variable*. [7].Contoh penggunaan *dummy variable* ditunjukkan pada gambar 2.1.



Gambar 2.1. Contoh penggunaan *dummy variable*

2.4. Penyakit Kandungan

Penyakit kandungan merupakan penyakit yang rentan diderita oleh setiap wanita. Penyakit kandungan menyerang organ genital (organ reproduksi) wanita yang dapat menyebabkan kemandulan. Jenis penyakit ini cukup banyak, beberapa di antaranya adalah radang panggul, mioma uteri, kanker serviks, dan kanker ovarium. keempat jenis penyakit tersebut menyerang organ genital internal atau bagian dalam yang terdiri dari

ovarium, tuba fallopi, uterus, endomerium, serviks, dan vagina.

a. Radang Panggul

Radang panggul atau *pelvic inflammatory disease* (PID) adalah infeksi traktus genital atas yang merupakan salah satu komplikasi dari infeksi menular seksual (IMS) [8]. Gejala penyakit radang panggul berupa: nyeri perut bagian bawah [9], temperatur oral lebih dari 38,3°C [9], keluar cairan dari vagina [9], pendarahan tidak teratur [10], sakit kepala, lesu [9], nyeri berhubungan seksual [8] dan nyeri buang air kecil [8].

b. Mioma Uteri

Mioma Uteri adalah tumor jinak otot polos yang terdiri dari sel-sel jaringan otot polos, jaringan fibroid dan kolagen [11]. Mioma sangat bergantung pada hormon esterogen. Mioma dianggap cukup tidak berbahaya namun dapat menimbulkan masalah pada saat persalinan [12]. Gejala mioma uteri adalah sebagai berikut: frekuensi buang air kecil bertambah [10], sulit buang air besar [9], pendarahan mens abnormal [11], terdapat benjolan di perut bagian bawah [9], pendarahan diluar siklus haid [11], nyeri haid [9], nyeri ketika berhubungan seksual [10], anemia [11], nyeri panggul [11], nyeri punggung [11], dan infertilitas [9].

c. Kanker Serviks

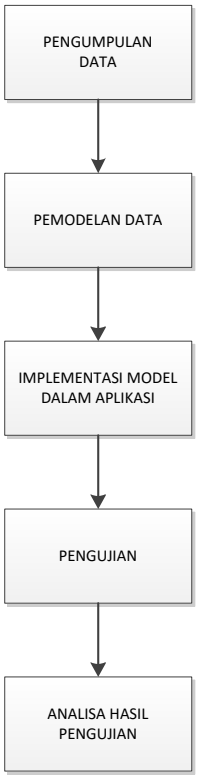
Kanker serviks merupakan kelainan yang terjadi pada sel-sel serviks yang berkembang dengan cepat dan tidak terkontrol [13]. Kanker serviks 80% disebabkan oleh HPV (*Human Papilloma Virus*) [10]. Gejala kanker serviks meliputi: pendarahan abnormal [10], pendarahan haid abnormal [13], nyeri panggul [13], nyeri ketika berhubungan seksual [13], keputihan [13], nyeri buang air kecil [14], anemia [10], nyeri punggung bagian bawah [14], penurunan nafsu makan [14] dan penurunan berat badan drastis [14].

d. Kanker Ovarium

Kanker ovarium adalah kanker yang bermula pada indung telur (ovarium) wanita. Kanker ovarium memiliki potensi menyebarkan sel ganas dengan sangat cepat ke seluruh rongga abdomen [12]. Gejala kanker ovarium adalah sebagai berikut: nyeri ketika berhubungan seksual [15], kembung [14], sulit buang air besar [9], sering buang air kecil [15], nafsu makan menurun [14], cepat lelah, anemia [10], nyeri panggul [15] dan nyeri punggung bagian bawah [10].

3. METODOLOGI PENELITIAN

Pelaksanaan penelitian ini dibagi menjadi 5 bagian, yaitu tahap pengumpulan data, tahap analisa dan perancangan, tahap implementasi, tahap pengujian, dan tahap analisa hasil pengujian. Gambar 3.1 menunjukkan tahapan penelitian.



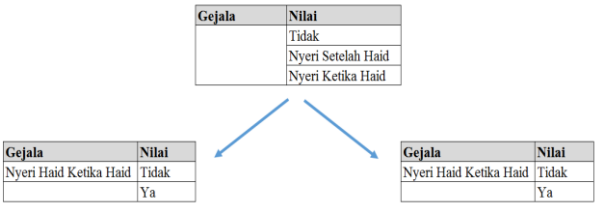
Gambar 3.1. Tahapan metodologi penelitian

3.1. Pengumpulan Data

Pengumpulan data dilakukan dengan mengambil data pada penelitian sebelumnya oleh [3] yang merupakan data rekam medik pasien RSUD Dr. Moewardi Solo. Data yang digunakan pada penelitian ini adalah 125 data dengan 5 kelas penyakit dan 18 gejala penyakit. Rincian gejala penyakit dapat dilihan pada tabel 3.1 dan data kelas penyakit dapat dilihat pada tabel 3.2.

3.2. Pemodelan Data

Pemodelan data adalah membuat model pada data yaitu membuat data menjadi dua jenis yaitu data asli (data yang belum diaplikasikan dengan *Dummy Variable*) dan data yang sudah di aplikasikan dengan *dummy variable*, selanjutnya kedua data tersebut akan klasifikasikan dengan metode *C4.5*. Contoh hasil aplikasi *Dummy Variable* pada gejala penyakit kandungan dapat dilihat dari gambar 3.2.



Gambar 3.2. Aplikasi *dummy variable* pada gejala penyakit

3.3. Implementasi Model Dalam Aplikasi

Implementasi dilakukan dengan bahasa pemrograman Java menggunakan aplikasi pendukung sebagai berikut : Netbeans 8.0.1 untuk *editor*, dan API Weka 3.8.0.

3.4. Pengujian

Pengujian yang dilakukan pada penelitian ini menggunakan metode *k-fold cross validation.* Data penyakit kandungan yang digunakan untuk proses *training* hingga mendapatkan pohon keputusan dievaluasi dengan metode *k-fold cross validation* dengan nilai *k* = 5, 10, 15, 20, dan 25. Data pada metode ini dibagi menjadi *k* bagian secara acak, kemudian dilakukan *k* kali eksperimen dimana masing-masing eksperimen menggunakan bagian data ke-*k* sebagai data testing dan memanfaatkan bagian lainnya sebagai data training. Penghitungan akurasi hasil klasifikasi digunakan *precision, recall,* dan *f-measure*. Hasil klasifikasi disajikan menggunakan *confusion matrix* dengan nilai L=5 karena data diklasifikasi dalam 5 kelas. Tabel 3.1 menunjukkan penggunaan *confusion matrix*.

Tabel 3.1. *Confusion matrix* hasil klasifikasi

Kelas Sebenarnya	Kelas hasil klasifikasi					
	A	B	C	D	E	TOTAL
A	TpA	Error	Error	Error	Error	TotalA
B	Error	TpB	Error	Error	Error	TotalB
C	Error	Error	TpC	Error	Error	TotalC
D	Error	Error	Error	TpD	Error	TotalD
E	Error	Error	Error	Error	TpE	TotalE
TOTAL	Terprediksi_A	Terprediksi_B	Terprediksi_C	Terprediksi_D	Terprediksi_E	

Keterangan:

A = Kelas Tidak sakit

B = Kelas Radang Panggul

C = Kelas Mioma Uteri

D = Kelas Kanker Serviks

E = Kelas Kanker Ovarium

TP atau *True Positive* menunjukkan data yang diklasifikasi sesuai dengan kelas sebenarnya, sedangkan *Error* menunjukkan data yang diklasifikasikan tidak sesuai dengan kelas sebenarnya. Berdasarkan *confusion matrix* kemudian dilakukan penghitungan *accuracy, precision*, *recall*, dan *f-measure*. Berikut rumus penghitungan *accuracy*, *precision*, *recall*, dan *f-measure.*

$$Accuracy = \frac{TP\ (A+B+C+D+E)}{Total\ (A+B+C+D+E)} \times 100\% \quad (3.1)$$

$$Precission\_x = \frac{TP\_x}{Terprediksi\_x} \quad (3.2)$$

$$Recall\_x = \frac{TP\_x}{Total\_x} \quad (3.3)$$

$$F\_Measure\_x = \frac{2P\_x \times R\_x}{P\_x+R\_x} \quad (3.4)$$

Nilai *x* menunjukkan kelas yang dihitung yaitu A (Tidak sakit), B (Radang Panggul), C (Mioma Uteri), D (Kanker Serviks) dan E (Kanker Ovarium). Proses selanjutnya adalah menghitung rata-rata dari masing kelas untuk mendapatkan nilai *precision*, *recall*, dan *f-measure*.

3.5. Analisa Hasil Pengujian

Analisa hasil pengujian merupakan hasil evaluasi metode dan pengujian yang telah dilakukan sebelumnya.

4. HASIL DAN PEMBAHASAN

4.1. Deskripsi Data

Data penelitian ini merupakan data pada penelitian sebelumnya oleh [3]. Data pada penelitian ini mengandung 5 kelas penyakit dengan 18 gejala penyakit.. Keterangan tentang 5 kelas penyakit dan 18 gejala penyakit dapat dilihat pada tabel 4.1 dan 4.2.

Tabel 4. 1. Daftar kelas penyakit

No	Kode	Kategori Penyakit	Deskripsi
1	A	Tidak Sakit	Pasien tidak menderita salah satu dari 4 penyakit kelamin
2	B	Radang Panggul	Penyakit infeksi traktus genital atas yang merupakan salah satu komplikasi dari infeksi menular seksual [8]
3	C	Mioma Uteri	Tumor jinak otot polos yang terdiri dari sel-sel jaringan otot polos, jaringan fibroid dan kolagen [11]
4	D	Kanker Serviks	Merupakan kelainan yang terjadi pada sel-sel serviks yang berkembang dengan cepat dan tidak terkontrol [13]
5	E	Kanker Ovarium	Kanker yang bermula pada indung telur (ovarium) wanita. [12]

Tabel 4. 2. Daftar gejala penyakit

No	Gejala	Deskripsi	Nilai
1	Anemia	keadaan saat jumlah sel darah merah berada di bawah normal karena pendarahan [11].	(Tidak, Ya)
2	Nyeri Haid	Nyeri yang dirasakan pada bagian perut bawah pada saat masa menstruasi [9]	(Tidak, Nyeri setelah haid, Nyeri ketika haid)
3	Susah Hamil	Keadaan dimana wanita susah mengalami kehamilan dikarenakan gangguan transportasi sperma untuk pembuahan sel telur [11].	(Tidak, Ya)
4	Benjolan Perut	Munculnya benjolan di bagian perut atas atau bawah [9].	(Tidak, Ya)
5	Pendarahan	Pendarahan yang terjadi pada uterus wanita [11].	(Tidak, Pendarahan Menstruasi Abnormal, Pendarahan Tiba-tiba)
6	Nyeri Berhubungan Seksual	Nyeri yang terjadi saat berhubungan seksual karena adanya penekanan tumor pada daerah panggul [15].	(Tidak, Ya)
7	Cepat Lelah	Kondisi dimana kondisi badan cepat mengalami kelelahan dalam melakukan aktifitas sehari-hari [14].	(Tidak, Ya)
8	penurunan Berat Badan	Menurunnya berat badan karena nafsu makan menurun [14].	(Tidak, Ya)
9	Nyeri Panggul	Nyeri yang disebakan tumor yang membesar pada rongga pelvik yang menekan saraf [11].	(Tidak, Ya)
10	Gangguan Pencernaan	Gangguan pencernaan yang biasanya meliputi diare ataupun sembelit [9]	(Tidak, Ya)
11	Nyeri Perut	Nyeri yang tejadi area perut [9]	(Tidak, Nyeri pada rongga perut, Nyeri perut bagian bawah, Nyeri perut bagian pinggul)
12	Nyeri Punggung	Nyeri yang disebakan tumor yang membesar pada rongga pelvik yang menekan saraf sampai bagian punggung [11].	(Tidak, Ya)
13	Penurunan Nafsu Makan	Penurunan nafsu makan akibat perut terasa penuh, mual, dan kembung [14].	(Tidak, Ya)
14	Demam	Keadaan dimana temperatur diatas 38.3 derajat celcius [9].	(Tidak, Ya)
15	Sakit Kepala	Kondisi dimana kepala mengalami nyeri [9].	(Tidak, Ya)
16	Kembung	Kondisi dimana perut terasa penuh dan kencang [14].	(Tidak, Ya)
17	Keputihan	Keluarnya cairan bukan darah yang keluar melalui organ genital wanita [13]	(Tidak, Ya)
18	Gangguan BAK	Gangguan buang air kecil karena penekanan terhadap kandung kemih oleh tumor [15].	(Tidak, Sering BAK, Nyeri BAK, Nyeri dan Sering BAK,

Tabel 4. 3. Daftar gejala penyakit (lanjutan)

4.2. Pemodelan Data

Data pada tahap ini dimodelkan menjadi 2 jenis yaitu data asli (data tanpa aplikasi *dummy variable*) dan data yang sudah diaplikasikan *dummy variable.* Hasil transformasi gejala penyakit kandungan setelah aplikasi *dummy variable* dapat terlihat di tabel 4.3.

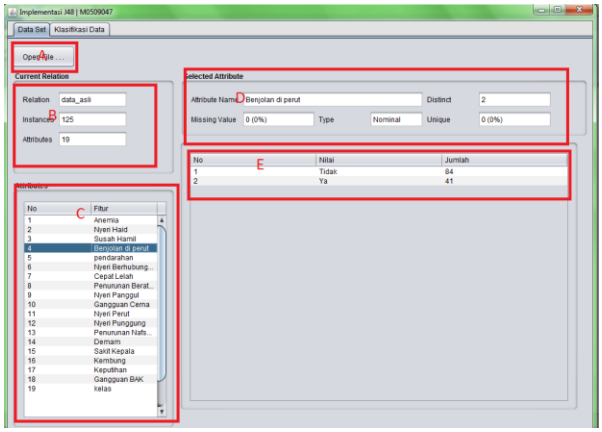
Tabel 4. 4. Hasil aplikasi Dummy Variable pada gejala

No	Gejala	Nilai
1	Anemia	(Tidak , Ya)
2	Nyeri Haid_Setelah Haid	(Tidak , Ya)
3	Nyeri Haid_Ketika Haid	(Tidak , Ya)
4	Susah Hamil	(Tidak , Ya)
5	Benjolan Perut	(Tidak , Ya)
6	Pendarahan_Menstruasi_abnormal	(Tidak , Ya)
7	Pendarahan_Tiba-tiba	(Tidak , Ya)
8	Nyeri Berhubungan Seksual	(Tidak , Ya)
9	Cepat Lelah	(Tidak , Ya)
10	penurunan Berat Badan	(Tidak , Ya)
11	Nyeri Panggul	(Tidak , Ya)
12	Gangguan Pencernaan_sembelit	(Tidak , Ya)
13	Gangguan Pencernaan_diare	(Tidak , Ya)
14	Nyeri Perut_Rongga Perut	(Tidak , Ya)
15	Nyeri Perut_Bagian Bawah	(Tidak , Ya)
16	Nyeri Perut_ Pinggul	(Tidak , Ya)
17	Nyeri Punggung	(Tidak , Ya)
18	Penurunan Nafsu Makan	(Tidak , Ya)
19	Demam	(Tidak , Ya)
20	Sakit Kepala	(Tidak , Ya)
21	Kembung	(Tidak , Ya)
22	Keputihan	(Tidak , Ya)
23	Gangguan BAK_Sering	(Tidak , Ya)
24	Gangguan BAK_Nyeri	(Tidak , Ya)

4.3. Implementasi

Sistem yang dibangun hanya digunakan sebagai pendukung, bukan sebagai hal yang difokuskan pada penelitian ini. Maka dari itu penulis tidak menjelaskan secara detail bagaimana alur dalam sistem ini. Pada tahap ini penulis melakukan pengembangan sistem yang dapat melakukan klasifikasi menggunakan metode *C4.5,* dan melakukan serta menampilkan hasil pengujian dengan metode *k-fold cross validation.*

Tampilan awal sistem yang dibangun dapat dilihat pada gambar 4.1.

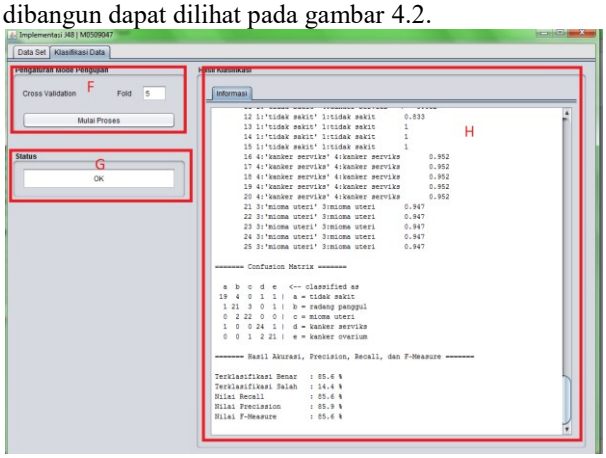


Gambar 4.1. Tampilan awal sistem

Keterangan:

A. Tombol untuk memilih file
B. Detail file
C. Daftar atribut
D. Detail selected atribut
E. Detail nilai selected atribut

Tampilan klasifikasi dan hasil pengujian yang dibangun dapat dilihat pada gambar 4.2.



Gambar 4.2. Tampilan klasifikasi dan hasil pengujian

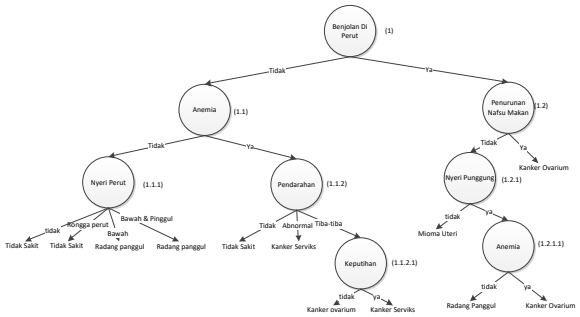
Keterangan:

F. Mode pengujian k-fold cross validation
G. Status sistem
H. Output klasifkasi dan pengujian

4.4. Pengujian

Pengujian menggunakan sistem yang telah dibuat untuk menghasilkan output klasifikasi berupa: informasi data, pohon keputusan C4.5, percobaan prediksi, confusion matrix, dan hasil *accuracy, precission, recall,* dan *f-measure*. Gambar 4.3 dan 4.4 menunjukkan pohon keputusan yang dihasilkan oleh klasifikasi pada data asli dan data *dummy*. Setelah dilakukan implementasi dengan metode C4.5 pada data asli menghasilkan kesimpulan dari pohon keputusan bahwa gejala penyakit yang dominan pada penyakit kandungan sebanyak 7 gejala

yaitu: benjolan di perut, anemia, penurunan nafsu makan, nyeri perut, pendarahan, nyeri punggung, dan keputihan.



Gambar 4.3. Pohon keputusan data asli

Sedangkan untuk data *dummy* menghasilkan kesimpulan bahwa gejala penyakit yang dominan pada penyakit kandungan sesuai pohon keputusan yang dihasilkan adalah sebanyak 10 gejala yaitu: nyeri perut bawah, pendarahan tiba-tiba, anemia, benjolan di perut, keputihan, demam, penurunan nafsu makan, susah hamil, nyeri panggul, dan gangguan BAK nyeri.

Gambar 4.4. Pohon keputusan data *dummy*

Pada hasil pengujian pohon keputusan diperoleh atribut yang dominan dari kedua pohon keputusan. Tabel 4.4 menunjukkan atribut gejala yang dominan dari data asli dan data *dummy*.

Tabel 4.4. Atribut gejala yang dominan

Data Asli	Data *Dummy*
Benjolan di perut (1)	Benjolan di perut (1.2.1)
Anemia (1.1)	Anemia (1.2)
Penurunan nafsu makan (1.2)	Penurunan nafsu makan (1.2.2.1)
Keputihan (1.1.2.1)	Keputihan (1.2.2)
Nyeri perut (1.1.1)	Nyeri perut bawah (1)
Pendarahan (1.1.2)	Pendarahan tiba – tiba (1.1)
Nyeri punggung (1.2.1)	
	Demam (1.2.1.1)
	Susah hamil (1.2.1.1.1)
	Nyeri panggul (1.2.2.1.1)
	Gangguan BAK Nyeri (1.2.2.2.1)

Selanjutnya pengujian yang dilakukan adalah untuk mengetahui nilai *accuracy, precission, recall,* dan *f-measure* menggunakan metode *k-fold cross validation.* Data asli dan data hasil *dummy* diuji beberapa kali menggunakan metode *k-fold cross validation* dengan nilai k = 5, 10, 15, 20, 25. Hal ini dilakukan bukan hanya karena nilai *k* sebagai nilai pembagi dari data yang berjumlah 125 tapi juga nilai kelipatan untuk jumlah kelas yang terdapat pada data penyakit kandungan.

Pengujian pertama dilakukan untuk data asli. Berikut hasil pengujian data asli pada *5-fold cross validation*:

Tabel 4.5. *Confusion matrix* data asli pada *5-fold cross valdiation*

Kelas sebenarnya	Kelas prediksi					
	A	B	C	D	E	TOTAL
A	19	4	0	1	1	25
B	1	21	3	0	1	26
C	0	2	22	0	0	24
D	1	0	0	24	1	26
E	0	0	1	2	21	24
TOTAL	21	27	26	27	24	

$Pa = \frac{19}{19+1+1} = 0.905$
$Pb = \frac{21}{21+4+2} = 0.778$
$Pc = \frac{22}{22+3+1} = 0.846$
$Pd = \frac{24}{24+1+2} = 0.889$
$Pe = \frac{21}{21+1+1+1} = 0.875$

$Ra = \frac{19}{19+4+1+1} = 0.760$
$Rb = \frac{21}{21+1+3+1} = 0.808$
$Rc = \frac{22}{22+2} = 0.917$
$Rd = \frac{24}{24+1+1} = 0.923$
$Re = \frac{21}{21+1+2} = 0.87.5$

$Fa = \frac{2Pa\ x\ Ra}{Pa+Ra} = 0.826$
$Fb = \frac{2Pb\ x\ Rb}{Pb+Rb} = 0.792$
$Fc = \frac{2Pc\ x\ Rc}{Pc+Rc} = 0.880$
$Fd = \frac{2Pd\ x\ Rd}{Pd+Rd} = 0.906$
$Fe = \frac{2Pe\ x\ Re}{Pe+Re} = 0.875$

Precission $= \frac{Pa+Pb+Pc+Pd+Pe}{5} x\ 100\% = 85.8\ \%$

Recall $= \frac{Ra+Rb+Rc+Rd+Re}{5} x\ 100\% = 85.65\ \%$

F – Measure $= \frac{Fa+Fb+Fc+Fd+Fe}{5} x\ 100\% = 85.58\ \%$

Accuracy $= \frac{Total\ Terklasifikasi\ Benar}{Banyaknya\ Data} = \frac{107}{125} x\ 100\% = 85.6\ \%$

Pengujian kedua dilakukan untuk data dummy. Berikut hasil pengujian data dummy pada 5-fold cross validation:

Tabel 4.6. *Confusion matrix* data dummy pada 5-fold cross valdiation

Kelas sebenarnya	Kelas prediksi					
	A	B	C	D	E	TOTAL
A	22	0	0	2	1	25
B	1	23	2	0	0	26
C	0	2	20	0	2	24
D	2	0	1	23	0	26
E	0	0	1	2	21	24
TOTAL	25	25	24	27	24	

$Pa = \frac{22}{22+1+2} = 0.880$
$Pb = \frac{23}{23+2} = 0.920$
$Pc = \frac{20}{20+2+1+1} = 0.833$
$Pd = \frac{23}{23+2+2} = 0.852$
$Pe = \frac{21}{21+1+2} = 0.875$

$Ra = \frac{22}{22+2+1} = 0.880$
$Rb = \frac{23}{23+1+2} = 0.885$
$Rc = \frac{20}{20+2+2} = 0.833$
$Rd = \frac{23}{23+2+1} = 0.885$
$Re = \frac{21}{21+1+2} = 0.875$

$Fa = \frac{2Pa\ x\ Ra}{Pa+Ra} = 88.0\ \%$
$Fb = \frac{2Pb\ x\ Rb}{Pb+Rb} = 90.2\ \%$
$Fc = \frac{2Pc\ x\ Rc}{Pc+Rc} = 83.3\ \%$
$Fd = \frac{2Pd\ x\ Rd}{Pd+Rd} = 86.8\ \%$
$Fe = \frac{2Pe\ x\ Re}{Pe+Re} = 87.5\ \%$

Precission $= \frac{Pa+Pb+Pc+Pd+Pe}{5}\ x\ 100\% = 87.2\ \%$

Recall $= \frac{Ra+Rb+Rc+Rd+Re}{5}\ x\ 100\% = 87.2\ \%$

F – Measure $= \frac{Fa+Fb+Fc+Fd+Fe}{5}\ x\ 100\% = 87.2\ \%$

Accuracy $= \frac{Total\ Terklasifikasi\ Benar}{Banyaknya\ Data} = \frac{109}{125}\ x\ 100\% = 87.2\ \%$

Hasil dari seluruh pengujian pada data asli dan data *dummy* dapat dilihat pada tabel 4.7.

Tabel 4.7. Hasil pengujian *k-fold cross validation* pada data asli dan data *dummy*

No	fold	Accuracy (%)		Precission (%)		Recall (%)		F-measure (%)	
		Asli	Dummy	Asli	Dummy	Asli	Dummy	Asli	Dummy
1	5	85.6	87.2	85.9	87.2	85.7	87.2	85.6	87.2
2	10	92.8	87.2	92.8	87.2	92.9	87.2	92.8	87.2
3	15	89.6	87.2	89.7	87.2	89.7	87.2	89.6	87.2
4	20	90.4	87.2	90.6	87.2	90.5	87.2	90.4	87.2
5	25	89.6	87.2	89.7	87.2	89.7	87.2	89.6	87.2
Mean		**89.6**	**87.2**	**89.74**	**87.2**	**89.7**	**87.2**	**89.6**	**87.2**

4.5. Analisa Hasil Pengujian

Setelah dilakukan implementasi dengan metode C4.5 pada data asli dan data *dummy*, muncul atribut yang dominan berdasarkan hasil pohon keputusan C4.5. Untuk data *dummy* muncul atribut yang lebih spesifik dibandingkan dengan data asli seperti, atribut nyeri perut bawah pada data *dummy* dengan atribut nyeri perut pada data asli dan atribut pendarahan tiba – tiba pada data *dummy* dengan atribut pendarahan pada data asli. Hal ini menyatakan bahwa penggunaan *dummy variable* dapat menspesifikasikan atribut gejala penyakit kandungan yang digunakan untuk menjadi parameter gejala penyakit kandungan.

Untuk pengujian *accuracy, precission, recall,* dan *f-measure* dilakukan dengan menggunakan lima nilai *fold* yang berbeda pada pengujian *k-fold cross validation* dapat disimpulkan bahwa penggunaan *dummy variable* pada klasifikasi penyakit kandungan dengan metode C4.5 menghasilkan nilai yang konsisten untuk kelima *fold* pada pengujian *k-fold cross validation*. Namun, penggunaan *dummy variable* pada klasifikasi penyakit kandungan dengan metode C4.5 menghasilkan nilai rata – rata *accuracy, precission, recall,* dan *f-measure* yang lebih rendah daripada data asli yang tanpa menggunakan *dummy variable*. Akan tetapi, untuk pengujian pada *fold* 5 pada data *dummy* menghasilkan nilai *accuracy, precission, recall,* dan *f-measure* yang lebih tinggi daripada data asli.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Dari hasil penelitian dapat disimpulkan bahwa penggunaan *dummy variable* menghasilkan nilai rata-rata *accuracy, precission, recall,* dan *f-measure* yang stabil secara berurutan yaitu sebesar 87.2%, 87.2%, 87.2%, dan 87.2% pada pengujian *k-fold cross validation* dengan nilai *fold* (5, 10, 15, 20, dan 25). Akan tetapi, penggunaan dummy variable mengurangi nilai rata – rata *accuracy, precission, recall,* dan *f-measure* secara berurutan dari 89.6%, 89.74%, 89.7%, dan 89.6% menjadi 87.2%, 87.2%, 87.2%, dan 87.2%.

Selain itu penggunaan *dummy variable* menspesifikasikan atribut gejala penyakit yang digunakan pada klasifikasi penyakit kandungan hal ini ditunjukkan dengan munculnya atribut yang dominan berdasarkan hasil pohon keputusan C4.5. Untuk data *dummy* muncul atribut yang lebih spesifik dibandingkan dengan data asli seperti, atribut nyeri perut bawah pada data *dummy* dengan atribut nyeri perut pada data asli dan atribut pendarahan tiba – tiba pada data *dummy* dengan atribut pendarahan pada data asli.

5.2. Saran

Saran untuk penelitian selanjutnya adalah melanjutkan penelitian pengaruh penggunaan *dummy variable* pada klasifikasi penyakit kandungan dengan menggunakan metode *decision tree* selain *C4.5* seperti BF Tree, NB Tree, Random Tree dan, Simple Cart seperti yang dilakukan oleh [16] , karena pada penelitian ini beberapa metode tersebut menghasilkan tingkat akurasi yang sama dengan metode C4.5. Sehingga perlu dilakukan penelitian terhadap metode tersebut untuk mengetahui apakah pengaruh penggunaan *dummy variable* dibandingkan dengan metode C4.5.

6. DAFTAR PUSTAKA

[1] A. Oluwapelumi, "Incorporating Dummy Variables In Regression Model To Determine The Average Internally Generated Revenue And Wage Bills Of The Six Geopolitical Zone In Nigeria," *European Journal of Statistic and Probability. Vol.2, No. 1,* pp. 23-27, 2014.

[2] P. A. Nugraha, "Perbandingan Metode Probablilistic Naïve Bayes Classifier Dan Jaringan Syaraf Tiruan Learning Vector Quantization Dalam Kasus Klasifikasi Penyakit Kandungan," Surakarta, 2014.

[3] A. Prabhawaningrum, "Perbandingan Algoritma Levenberg-Marquadt Dengan Backpropagation Untuk Mendiagnosa Jenis Penyakit Kandungan," Surakarta, 2013.

[4] S. Kharya, "Using Data Mining Techniques for Diagnosis And Prognosis of Cancer Disease," *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2,* 2012.

[5] R. D. &. S. Y. Sari, "Penerapan Data Mining Untuk Analisa Pola Perilaku," *Jurnal ilmiah Teknologi dan Informasi ASIA,* vol. 8, p. 10, 2014.

[6] I. H. F. E. &. H. M. A. Witten, Data Mining : Practical Machine Learning Tools and Techniques (3rd ed), Morgan Kauffman, 2011.

[7] D. N. Gujarati, Basic Econometrics, McGraw Hill. p. 1002. ISBN 0-07-0233542-4, 2003.

[8] F. Thjay, "Ancaman Penyakit Radang Panggul Pada Infeksi Menular Seksual," *Majalah Kedokteran Damianus,* vol. 8, pp. 105-114, 2009.

[9] T. Nugroho, Buku Ajar Ginekologi Untuk Mahasiswa Kebidanan, Yogyakarta: Nuhamedia, 2010.

[10] E. &. S. J. Norwitz, At Glance Obstetri & Ginekologi, Jakarta: Erlangga, 2008.

[11] B. R. Hardibroto, "Mioma Uteri," *Majalah Kedokteran Nusantara,* vol. 38, pp. 254 - 259, 2005.

[12] C. &. T. E. Livoti, Menyingkap Tabir Yang Selama Ini Tersembunyi Tentang Vagina, Jakarta: Indeks, 2006.

[13] Mayangsari, 2010. [Online]. Available: http://angsamerah.com/img/Kanker Serviks.pdf. [Accessed 30 January 2013].

[14] R. E. Harahap, Kanker Ginekologi, Jakarta: Gramedia, 1984.

[15] P. A. A. &. L. W. Widayati, "Produksi Kit Imunoradiometricassay (IRMA) CA - 125 Untuk Deteksi Dini Kanker Ovarium," *Jurnal Ilmu Kefarmasian,* vol. 7, pp. 91-97, 2009.

[16] S. G. G. K. S. Ozsoy, "C4.5 Versus Other Decision Trees: A Review," *Computer Engineering and Applications,* vol. 04, September 2015.

[17] S. S. S. Garavaglia, "A Smart Guide To Dummy Variables : Four Application And Macro," Muray Hill, New Jersey, 1998.