Jurnal Informatika, Vol. 12, No. 2, November 2014, 70-76 DOI: 10.9744/informatika.12.2.70-76
ISSN 1411-0105

IDENTIFIKASI FAKTOR PREDIKSI DIAGNOSIS TINGKAT KEGANASAN KANKER PAYUDARA METODE STEPWISE BINARY LOGISTIC REGRESSION

Retno Aulia Vinarti[1*], Wiwik Anggraeni[1]
[1]Jurusan Sistem Informasi, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember
Jl. Raya ITS, Surabaya, 60111, Telp: (031) 5964965, Fax: (031) 5999944
E-mail: vaulia@gmail.com
*Korespondensi penulis

Abstrak: Organisasi Kesehatan Dunia (WHO) melaporkan bahwa kasus kematian di dunia akibat kanker empat tahun terakhir telah meningkat cukup tajam. Data peningkatan tersebut juga terjadi pada kasus kanker payudara. Di Indonesia sendiri, dua kasus ini juga merupakan kasus tertinggi pembunuh wanita. Berdasarkan Sistem Informasi Rumah Sakit (SIMRS), jumlah pasien kanker payudara baik rawat inap maupun rawat jalan adalah sebesar 28.7%. Fakta yang terungkap yaitu lebih dari 40% dari semua kanker dapat dicegah dengan catatan bahwa kanker harus terdeteksi lebih dini. Peran Teknologi Informasi dapat dirupakan dengan teknik penggalian data untuk mempersingkat waktu, akurasi dan pemilihan faktor pendeteksian dini penyakit kanker payudara. Metode stepwise binary logistic regression memiliki keunggulan untuk menambah dan mengurangi variabel independen sesuai dengan tingkat signifikansi dari model yang terbentuk. Berdasarkan hasil analisis pembobotan, empat variabel tertinggi yang harus lebih diwaspadai adalah luas area dari kanker (area), tingkat kehalusan (smoothness), banyaknya titik (concave points) atau inti kanker dan tingkat keabu-abuan dari kanker (texture). Sehingga akurasi dan kecepatan pemrosesan dari diagnosis tingkat keparahan kanker payudara dapat ditingkatkan melalui metode ini. Terlebih apabila metode ini dapat mengurangi jumlah kematian yang disebabkan oleh terlambatnya diagnosis tingkat keparahan kanker payudara.

Kata kunci: kanker payudara, prediksi, regresi, tingkat keparahan

Abstract:
The World Health Organization (WHO) reported that deaths caused by cancer in the world these last four years has increased significantly. The data also reflected in the increase in breast cancer cases. In Indonesia, two cases also the highest cases of adult female deaths. Based on Hospital Information System, the number of breast cancer patients either inpatient or outpatient care amounted to 28.7%. This fact revealed more than 40% of all cancers can be prevented with early detection cancer. Role of Information Technology can implemented by data mining techniques to shorten the diagnosing time, accuracy and selection of factors early detection of breast cancer. Stepwise binary logistic regression method has the advantage to add and subtract the independent variables in accordance with level of significance of the model. Based on the analysis of weighting method, the highest four variables that should be more aware is the area of cancer (area), fineness (smoothness), the number of dots (concave points) or the nucleus of cancer and grayish level of cancer (texture). So the accuracy and processing speed of diagnosis of the severity of breast cancer can be improved through this method

Keywords*: Breast Cancer, prediction, regression, severity*

PENDAHULUAN

Kanker payudara dan kanker serviks merupakan dua penyakit tak menular yang sangat membahayakan tidak hanya kaum wanita, namun kini juga kaum pria. Organisasi Kesehatan Dunia (WHO) melaporkan bahwa kasus kematian di dunia akibat kanker empat tahun terakhir telah meningkat cukup tajam. Data peningkatan tersebut juga terjadi pada kasus kanker payudara. Sebanyak 1.7 juta wanita didiagnosis menderita penyakit ini pada tahun 2012. Angka ini juga merupakan angka tertinggi penyebab kematian bagi wanita di dunia (Febrida, 2013). Di Indonesia sendiri, dua kasus ini juga merupakan kasus tertinggi pembunuh wanita. Berdasarkan Sistem Informasi Rumah Sakit (SIMRS), jumlah pasien kanker payudara baik rawat inap maupun rawat jalan adalah sebesar 28.7%. Disusul pada peringkat kedua yaitu kanker serviks sebanyak 12.8% (D-13, 2014).

Menurut Direktur Pengendalian Penyakit Tidak Menular Kementrian Kesehatan, dr. Ekowati Rahajeng, permasalahan kanker di Indonesia sama dengan permasalahan yang dialami oleh kebanyakan negara berkembang lainnya. Tiga permasalahan utama

tersebut adalah kanker tidak dapat dideteksi, tidak dapat dicegah dan tidak dapat disembuhkan. Padahal fakta yang terungkap yaitu lebih dari 40% dari semua kanker dapat dicegah dengan catatan bahwa kanker harus terdeteksi lebih dini (D-13, 2014).

Salah satu peran Teknologi Informasi dalam pendeteksian dini telah diimplementasikan melalui aplikasi Spot it Yourself (Nawawi, 2013). Aplikasi ini memberikan informasi yang bersifat eksplanatoris mengenai cara untuk mendeteksi kanker payudara sejak dini. Peran Teknologi Informasi lainnya juga dapat dirupakan dengan teknologi penggalian data untuk mempersingkat waktu dan faktor pendeteksian dini penyakit kanker payudara. Data mengenai faktor-faktor diagnosis keganasan kanker payudara dapat diakses secara luas dan bebas (tidak berbayar) pada situs UCI Machine Learning. Sehingga peluang riset untuk akurasi dan deteksi dini sangat luas.

Penelitian sebelumnya yang terkait dengan deteksi dini penyakit kanker payudara telah dilakukan dengan cara komparasi berbagai algoritma klasifikasi. Penelitian ini bertujuan untuk mempertajam hasil analisis sebelum proses diagnosa menggunakan beragam algoritma klasifikasi. Ketajaman yang dimaksud adalah memilih dengan seksama variabel-variabel manakah yang benar-benar memiliki dampak yang relevan terhadap pendeteksian dini tingkat keganasan kanker payudara. Pemilihan variabel ini harapannya akan memiliki dampak terhadap akurasi dan terlebih waktu pemrosesan karena berkurangnya data yang diproses secara signifikan.

Beberapa teknik statistika dapat digunakan untuk pemilihan variabel sebelum pemrosesan klasifikasi. Salah satunya yang paling cocok dengan karakteristik data numerik sesuai dengan data Breast Cancer Wisconsin adalah teknik regresi. Pemilihan teknik regresi juga harus dilakukan dengan cermat. Berdasarkan fungsinya, stepwise regression adalah salah satu teknik regresi yang dapat memilih dengan cermat mana saja variabel-variabel independen yang akan mempengaruhi akurasi dari variabel dependen. Ini adalah kelebihan stepwise regression dibandingkan regresi biasa. Namun, regresi biasa pun memiliki beragam jenisnya. Berdasarkan sifat linearitas dari data, regresi memiliki tipe regresi linear dan non-linear. Dalam penelitian ini telah dilakukan uji coba awal bahwa regresi non-linear menunjukkan hasil yang jauh lebih baik dibandingkan dengan regresi linear. Tepatnya non-linear berjenis regresi logistik (Garson, 2011). Selain berdasarkan sifat linearitas dari data, regresi juga dibedakan berdasarkan jumlah dari variabel independen yang diikutsertakan. Regresi sederhana hanya mengikutsertakan satu variabel independen saja, sedangkan regresi berganda mencakup lebih dari satu variabel independen. Sesuai dengan data yang disediakan oleh situs UCI, variabel independen dapat dipastikan lebih dari satu. Hal ini menjadi landasan utama oleh peneliti untuk menggunakan persamaan regresi berganda. Satu faktor terakhir yang tidak kalah penting yaitu karakter dari variabel dependen. Hampir semua teknik penggalian data klasifikasi selalu memiliki kelas kategorikal dimana terdapat dua kemungkinan. Kemungkinan pertama yaitu terdiri dari dua kelas atau lebih dari dua kelas. Pada kasus Breast Cancer Wisconsin memiliki karakteristik variabel independen dua kelas. Dua kelas tersebut adalah kanker ganas (malignant) dan kanker jinak/tumor (benign). Oleh karena itu, pada penelitian ini lebih digunakan tipe regresi jenis binary dibandingkan dengan jenis multinomial.

DATA BREAST CANCER WISCONSIN

Data yang digunakan dalam penelitian ini adalah data Kanker Payudara Wisconsin. Pada situs UCI Machine Learning, data ini disediakan dalam dua varian yaitu prognosis dan diagnosis (Mangasarian, et al., 1992). Oleh karena tujuan dari penelitian ini adalah mengidentifikasi penyakit kanker payudara secara dini, sehingga data yang digunakan adalah data diagnosis. Dalam data ini terdiri dari 10 atribut penentu apakah kanker tersebut jinak atau ganas. Selain 10 atribut, data ini juga dilengkapi dengan nomor identifikasi pasien dan hasil diagnosis kanker. Sehingga keseluruhan dari data ini berjumlah total 12 atribut. Untuk setiap atribut, terdapat tiga jenis pengukuran yaitu rata-rata (*mean*), galat standar (*standard error*) dan terburuk (*worst*). Sepuluh atribut tersebut adalah jarak rata-rata dari titik pusat ke tepi (*radius*), nilai simpangan baku dari tingkat keabu-abuan (*texture*), keliling (*perimeter*), luas area (*area*), variasi lokal dari nilai radius (*smoothness*), nilai kuadrat dari keliling dibagi dengan luas area dikurangi 1 (*compactness*), tingkat keparahan dari kontur (*concavity*), jumlah konkaf (*concave points*), simetris (*symmetry*), perkiraan tepi garis dikurangi 1 (*fractal dimension*). Data ini memiliki 357 baris untuk diagnosis kanker jinak dan 212 baris untuk diagnosis kanker ganas.

Keseluruhan data memiliki skala ratio, kecuali untuk atribut kelas yaitu nominal (B/M). Untuk prapemrosesan data, ketigapuluh atribut ini akan dilabeli M, SE dan W. M merepresentasikan Mean, SE berarti data Standard Error dan W adalah data Worst. Sehingga M-Radius berarti radius untuk data rata-rata.

STEPWISE BINARY LINEAR REGRESSION

Untuk selanjutnya, ketigapuluh atribut tersebut akan diproses menggunakan regresi linear binary.

Regresi logistik adalah jenis regresi yang dapat digunakan untuk memprediksikan variabel dependen yang memiliki sifat kategorikal. Sedangkan untuk data independen dapat berupa skala numerikal continuous atau kategorikal. Regresi logistik dapat digunakan untuk menentikan besarnya dampak dari setiap variabel independen kepada variabel dependen. Regresi Logistik Binary atau binomial adalah bentuk regresi yang digunakan ketika variabel dependen bersifat dikotomi atau dua kelas. Bentuk regresi yang membolehkan lebih dari dua kelas variabel adalah Regresi Logistik Multinomial.

Cara kerja dari regresi logistik binary adalah mengimplementasikan estimasi kesamaan maksimal setelah mengubah variabel-variabel dependen menjadi variabel logit. Variabel logit (logit[θ(x)]) adalah natural log dari batas nilai apakah suatu variabel dependen dapat ditransformasi (θ) menjadi suatu nilai tertentu atau tidak. Biasanya, variabel logit pada persamaan regresi logistik binary adalah 1. Apabila pada persamaan regresi logistik multinomial adalah nilai tertinggi dari variabel dependen yang telah diberikan (Garson, 2011).

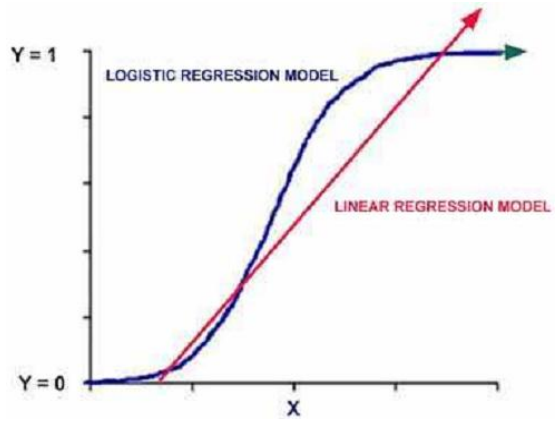
Persamaan θ dan logit[θ(x)] dapat dilihat pada persamaan berikut ini (Kyngas, et al., 2001)

$$\theta = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}$$

Dimana α adalah konstanta dari persamaan dan β adalah koefisien dari variabel independen (Agresti, 2002). Sedangkan persamaan alternatif dari regresi logistik adalah seperti persamaan berikut ini;

$$\text{logit}\left[\theta(\mathbf{x})\right] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

Gambar 1 menunjukkan perbedaan antara regresi linear dengan regresi logistik binary.



Gambar 1. Perbedaan model regresi logistik dengan regresi linear.

Regresi logistik stepwise tersedia baik dalam regresi logistik binary, multinomial maupun regresi linear biasa. Metode stepwise memiliki keunggulan untuk menambah dan mengurangi variabel independen sesuai dengan tingkat signifikansi dari model yang terbentuk. Tentu saja, untuk melakukan hal ini diperlukan sejumlah iterasi yang akan memroses dan menganalisis setiap model baru yang terbentuk karena pertambahan atau pengurangan variabel dependen.

Untuk menilai seberapa bagus persamaan regresi logistik binary yang terbentuk, stepwise memiliki beberapa cara untuk menguji. Uji tersebut antara lain maximum likelihood yang dinilai dengan likelihood ratio, hosmer-lemeshow goodness of fit dan Wald. Uji Wald adalah salah satu cara untuk menguji signifikansi dari sebagian variabel independen pada suatu model statistika. Dalam hal regresi logistik binary, variabel dependen akan bernilai 0 atau 1. Apabila suatu hasil uji test Wald signifikansinya kurang dari tingkat kepercayaan atau bernilai 0 maka variabel tersebut akan dimasukkan dalam model regresi logistik (Agresti, 1990). Uji Wald menggunakan uji Z statistics dengan formula sebagai berikut

$$z = \frac{\hat{\beta}}{SE}$$

Dimana SE adalah nilai Galat Standar (Standard Error).

METODOLOGI PENELITIAN

Langkah-langkah yang dilakukan untuk menyelesaikan penelitian ini mengikuti langkah-langkah dari penggalian data secara umum. Langkah tersebut memiliki tiga langkah utama yang akan diturunkan menjadi beberapa langkah pendukung. Tiga langkah utama tersebut adalah praproses, proses dan post-proses. Langkah praproses terdiri dari dua langkah pendukung, yaitu labelisasi variabel dan mengatur skala pengukuran setiap variabel. Labelisasi variabel menghasilkan tiga puluh label yaitu seperti yang ditunjukkan oleh Tabel 1. Setelah itu semua variabel akan diberikan skala pengukuran. Tiga puluh variabel akan diberi skala Scale yang mencakup skala numerikal baik interval maupun rasio akan masuk dalam skala ini. Sedangkan nomor Rekam Medis yang bersifat unik dan hasil diagnosis diberikan skala nominal.

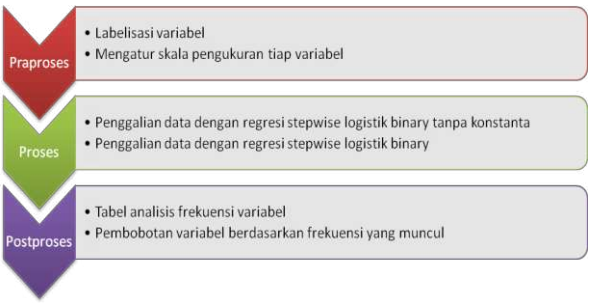
Tabel 1. Labelisasi, makna dan singkatan label

Label	Makna Label	Singkatan Label
No.RM	Nomor Rekam Medis	No.RM
Diagnosa	Hasil diagnosis: kanker jinak dan kanker ganas	Diag
MRadius	Data rata-rata untuk radius	MRad
MTexture	Data rata-rata untuk texture	MTex
MPerimeter	Data rata-rata untuk perimeter	MPer
MArea	Data rata-rata untuk area	MArea
MSmoothness	Data rata-rata untuk smoothness	MSmo
MCompactness	Data rata-rata untuk compactness	MComp
MConcavity	Data rata-rata untuk concavity	MConvt
MConcave	Data rata-rata untuk concave	MConv
MSymmetry	Data rata-rata untuk symmetry	MSym
MFractal	Data rata-rata untuk fractal	MFrac
SERadius	Data galat standar untuk radius	SERad
SETexture	Data galat standar untuk texture	SETex
SEPerimeter	Data galat standar untuk perimeter	SEPer
SEArea	Data galat standar untuk area	SEArea
SESmoothness	Data galat standar untuk smoothness	SESmo
SECompactness	Data galat standar untuk compactness	SEComp
SEConcavity	Data galat standar untuk concavity	SEConvt
SEConcave	Data galat standar untuk concave	SEConv
SESymmetry	Data galat standar untuk symmetry	SESym
SEFractal	Data galat standar untuk fractal	SEFrac
WRadius	Data terburuk untuk radius	WRad
WTexture	Data terburuk untuk texture	WTex
WPerimeter	Data terburuk untuk perimeter	WPer
WArea	Data terburuk untuk area	WArea
WSmoothness	Data terburuk untuk smoothness	WSmo
WCompactness	Data terburuk untuk compactness	WComp
WConcavity	Data terburuk untuk concavity	WConvt
WConcave	Data terburuk untuk concave	WConv
WSymmetry	Data terburuk untuk symmetry	WSym
WFractal	Data terburuk untuk fractal	WFrac

No.RM, Diagnosis, MRadius, MTexture, MPerimeter, MArea, MSmoothness, MCompactness, Mconcavity, MConcave, MSymmetry, MFractal, SERadius, SETexture, SEPerimeter, SEArea, SESmoothness, SECompactness, SEConcavity, SEConcave, SESymmetry, SEFractal, WRadius, Wtexture, WPerimeter, WArea, WSmoothness, Wcompactness, WConcavity, WConcave, WSymmetry, WFractal.

Langkah berikutnya yaitu langkah proses utama yaitu penggalian data dengan tujuan untuk diagnosis tingkat keganasan kanker payudara. Tingkat keganasan terdiri dari dua skala yaitu B atau M. B adalah Benign atau kanker jinak, sedangkan M adalah Malignant atau kanker ganas. Meskipun sebenarnya B dan M memiliki sifat ordinal atau ada tingkatan, namun karena hanya ada dua strata, maka hasilnya tidak akan berbeda antara dengan menggunakan regresi logistik binary atau ordinal. Proses utama memiliki dua proses pendukung yaitu penggalian data dengan regresi stepwise logistik binary tanpa konstanta dan dengan regresi stepwise logistik binary dengan konstanta. Perbedaan kedua metode ini terletak pada ada atau tidak nya konstanta yang diikutsertakan. Persamaan dari kedua persamaan ini adalah adanya faktor stepwise yang memungkinkan adanya proses eliminasi variabel prediktor secara otomatis.

Langkah terakhir yaitu postproses. Langkah ini bertujuan untuk membandingkan hasil akurasi dan efektivitas dari kedua persamaan regresi yang dihasilkan. Selain membandingkan hasil akurasi prediksi, langkah postproses juga melihat variabel manakah yang paling konsisten muncul di setiap persamaan regresi stepwise logistik binary. Tujuan dari analisis ini adalah untuk melihat tingkat urgensi dari variabel tersebut. Ilustrasi metodologi penelitian ini ditunjukkan pada gambar 2.



Gambar 2. Alur Metodologi Penelitian

HASIL dan PEMBAHASAN

Hasil penggalian data dengan regresi stepwise logistik binary tanpa konstanta dan dengan konstanta disajikan dalam dua tolak ukur yaitu jumlah langkah (step) dan persentase kebenaran dalam mendiagnosis tingkat keparahan kanker (Tabel 2).

Dari Tabel 1 dapat dilihat bahwa jumlah langkah dari metode regresi stepwise logistik binary tidak dipengaruhi baik dengan konstanta maupun tanpa konstanta. Tiga dari empat jenis data yang telah dilakukan penggalian data menujukkan bahwa persamaan regresi tanpa konstanta membutuhkan langkah yang lebih banyak. Sedangkan, hasil yang ditunjukkan oleh persamaan regresi logistik binary dengan konstanta menyatakan bahwa dengan berbekal data rata-rata dan galat standar mendapatkan persentase diagnosis yang lebih baik. Untuk data terburuk dan campuran menunjukkan performa diagnosis yang lebih baik bila digunakan oleh persamaan regresi logistik binary tanpa konstanta.

Peran konstanta pada suatu persamaan regresi diartikan sebagai faktor yang berasal dari luar variabel independen. Sehingga persamaan regresi logistik binary dengan konstanta membolehkan adanya faktor yang berasal dari selain kesepuluh variabel independen. Sedangkan pada persamaan regresi logistik binary tanpa konstanta mewajibkan persamaan garis yang terbentuk tanpa ada campur tangan dari faktor selain dari sepuluh variabel independen. Pada jaringan saraf tiruan, faktor konstanta ini mirip dengan peran dari bias.

Persamaan regresi logistik binary tanpa konstanta lebih ditujukan untuk mencari variabel independen manakah yang benar-benar mempengaruhi hasil diagnosis kanker payudara. Istilah yang sering digunakan untuk merepresentasikan kondisi ini adalah *feature selection*. Sedangkan persamaan regresi logistik binary dengan konstanta ditujukan untuk melakukan penggalian data sesuai dengan konstruksi dari persamaan regresi secara umum yang terdapat konstanta di dalamnya. Hasil persamaan regresi logistik binary ditunjukkan oleh Tabel 3.

Dari Tabel 2 dan Tabel 3 terlihat bahwa persamaan regresi yang paling akurat ditunjukkan oleh persamaan campuran tanpa konstanta. Persamaan tersebut memperoleh akurasi 98.4%. Pada posisi berikutnya masih diperoleh persamaan regresi tanpa konstanta namun menggunakan data *worst* sebesar 97.4%. Persamaan terakurat menggabungkan data *mean (Concave points, Symmetry, Fractal), SE (Area, Smoothness, Compactness), Worst (Texture, Concave, Symmetry, Fractal)*. Berdasarkan koefisien yang diperoleh, dapat diamati bahwa tiga besar variabel yang paling berpengaruh terhadap akurasi penentuan apakah kanker tersebut jinak atau ganas adalah MFrac, WFrac dan MConv. Oleh karena kedua persamaan ini tidak memiliki konstanta, maka semua penentuan tingkat keganasan kanker tidak mengikutsertakan faktor dari luar. Sedangkan hal yang sama juga terjadi pada persamaan regresi logistik dengan konstanta. Tiga variabel yang paling berpengaruh terhadap akurasi diagnosis secara berturut-turut adalah SEComp, WSmo, WConv. Namun persamaan ini juga memuat sebesar 52.65 bagian untuk faktor lain penyebab tingkat keganasan kanker payudara selain dari tiga puluh data yang disediakan.

Kedelapan persamaan regresi logistik pada Tabel 3 selanjutnya akan dianalisis frekuensi kemunculan dari setiap atribut. Penyusunan frekuensi kemunculan tersebut tidak melihat apakah data mean,

Tabel 2. Hasil penggalian data dengan metode regresi stepwise logistik binary

Jenis data	Jumlah langkah		Persentase mendiagnosa dengan tepat	
	Dengan konstanta	Tanpa konstanta	Dengan konstanta	Tanpa konstanta
Rata-rata (*mean*)	5	3	94.7%	93.1%
Galat standar (*SE*)	3	6	90.3%	89.1%
Terburuk (*worst*)	6	9	97.3%	97.4%
Campuran (*mix*)	**7**	**10**	**97.5%**	**98.4%**

Tabel 3. Persamaan Regresi Logistik Binary yang terbentuk

Jenis data	Persamaan Regresi Logistik Binary	
	Dengan konstanta	Tanpa konstanta
Rata-rata (*mean*)	Diag = 0.374 Mtex – 0.332 MPer + 0.035 MArea + 55.193 MSmo + 99.91 MConv – 9.264	Diag = 0.256 MTex + 143.018 MConv – 203.2 MFrac
Galat standar (*SE*)	Diag = -30.94 SERad + 0.429 SEArea + 15.729 SEConvt – 3.424	-27.53 SERad – 0.91 SETex + 0.376 SEArea – 158.4 SESmo + 20.339 SEConvt – 59.14 SESym
Terburuk (*worst*)	Diag = 0.277 WTex + 0.014 WArea + 49.781 WSmo + 36.96 WConv – 30.37	-3.557 WRad + 0.248 WTex + 0.049 WArea + 37.998 WSmo + 47.162 WConv
Campuran (*mix*)	Diag = 1.44 WRad + 0.371 WTex + 56.047 WSmo + 9.283 WConvt + 47.162 WConv + 15.57 SERad – 81.89 SEComp – 52.65	149.48 MConv – 49.41 MSym – 599.2 MFrac + 0.188 SEArea + 251.204 SESmo – 171.1 SEComp + 0.289 WTex + 14.192 WConv +33.575 WSym + 161.78 WFrac

SE atau Worst, melainkan dihitung dari kemunculan atribut data langsung. Hasil ringkasan ini ditunjukkan oleh Tabel 4. Tujuan dari penghitungan frekuensi ini adalah untuk melihat variabel manakah yang layak dijadikan acuan pertama kali oleh seorang ahli kanker untuk mendiagnosa tingkat keganasan dari kanker.

Terlihat pada Tabel 4 bahwa antara persamaan regresi dengan konstanta dan tanpa konstanta menunjukkan perbedaan. Oleh karena itu, proses analisis akan dilakukan secara terpisah sesuai dengan tujuan awal masing-masing persamaan.

Persamaan regresi logistik binary dengan konstanta yang bertujuan untuk melakukan penggalian data menunjukkan bahwa variabel yang paling sering muncul adalah luas area dari kanker (*area*), tingkat kehalusan (*smoothness*), banyaknya titik (*concave points*) atau inti kanker dan seberapa besar radius tiap titik menuju tepi (*radius*) dari kanker. Untuk data pendukung dapat dicari melalui variabel yang muncul dua kali yaitu tingkat keabu-abuan dari kanker (*texture*) dan tingkat keparahan dari kontur (*concavity*).

Sedangkan pada persamaan regresi logistik binary tanpa konstanta yang bertujuan untuk mengeliminasi faktor dari luar untuk proses diagnosis menunjukkan frekuensi yang berbeda. Terdapat empat kali kemunculan dari variabel texture dimana hal ini tidak ditemukan pada persamaan regresi logistik binary dengan konstanta. Selain itu, terdapat lima variabel yang juga berkontribusi terhadap justifikasi tingkat keparahan kanker payudara. Lima variabel tersebut adalah banyaknya inti kanker (*concave points*), perkiraan tepi garis dikurangi satu (*fractal dimension*), luas area kanker (*area*), tingkat kehalusan (*smoothness*) dan tingkat simetris (*symmetry*).

Untuk menentukan variabel manakah yang terpenting, maka diberikan bobot untuk setiap variabel berdasarkan urutan kemunculannya baik dalam persamaan regresi logistik binary dengan konstanta maupun tanpa konstanta. Sistem pembobotan dilakukan dengan cara memberikan bobot 4 apabila variabel tersebut 4 kali muncul, 3 apabila variabel tersebut 3 kali muncul. Setelah itu, ditambahkan antara bobot persamaan dengan konstanta dan tanpa konstanta lalu diurutkan mulai bobot paling tinggi hingga paling rendah. Hasil pembobotan dapat dilihat pada tabel 5 berikut.

Dari Tabel 5 dapat ditarik kesimpulan bahwa empat variabel yang memiliki bobot paling tinggi (6) adalah area, smoothness, concave points dan texture. Oleh karena itu, keempat variabel ini hendaknya menjadi rujukan pertama untuk melihat tingkat keparahan penyakit kanker payudara. Keempat variabel ini telah menjadi variabel yang terpilih berdasarkan metode stepwise binary logistic regression.

KESIMPULAN DAN SARAN

Kesimpulan dari penelitian ini akan ditarik dari berbagai sisi uji coba dan hasil yang didapatkan. Berdasarkan jumlah langkah atau iterasi yang harus ditempuh untuk mendapatkan hasil prediksi didapatkan bahwa persamaan data campuran (mix) selalu memiliki iterasi yang paling banyak. Hal ini wajar apabila melihat banyaknya data yang diproses pada data campuran tiga kali lipat dari data yang lain. Hal ini juga berimbang dengan akurasi yang diperoleh. Beragam tipe data membuat persamaan menjadi lebih kaya dan kompleks, oleh karena itu, akurasi tertinggi didapatkan oleh penggalian data campuran sebesar 98.4%.

Tabel 4. Hasil analisis frekuensi kemunculan atribut.

	Dengan konstanta	**Tanpa konstanta**
Empat kali muncul	-	Texture
Tiga kali muncul	Area, Smoothness, Concave, Radius	Concave, Fractal, Area, Smoothness, Symmetry
Dua kali muncul	Texture, Concavity	Radius, Concavity
Satu kali muncul	Perimeter, Compactness	Compactness
Tidak pernah muncul	Fractal, Symmetry	Perimeter

Tabel 5. Hasil pembobotan variabel

Variabel	**Bobot Persamaan Regresi dengan konstanta**	**Bobot Persamaan Regresi tanpa konstanta**	**Total Bobot**
Area	3	3	6
Smoothness	3	3	6
Concave Points	3	3	6
Radius	3	2	5
Texture	2	4	6
Concavity	2	2	4
Perimeter	1	0	1
Compactness	1	1	2
Fractal	0	3	3
Symmetry	0	3	3

Berdasarkan hasil persamaan regresi logistik binary yang dihasilkan, tiga besar variabel yang paling berpengaruh terhadap akurasi penentuan tingkat keparahan kanker payudara dalam persamaan regresi tanpa konstanta adalah MFrac, WFrac dan MConv. Sedangkan untuk persamaan regresi logistik binary dengan konstanta adalah SEComp, WSmo dan WConv. Namun persamaan ini juga memuat 52.65 bagian untuk faktor lain penyebab tingkat keganasan kanker payudara selain tiga puluh data yang disediakan.

Berdasarkan analisis pembobotan keseluruhan baik dari persamaan regresi dengan konstanta maupun tanpa konstanta didapatkan bahwa empat variabel tertinggi yang harus lebih diwaspadai adalah luas area dari kanker (area), tingkat kehalusan (smoothness), banyaknya titik (concave points) atau inti kanker dan tingkat keabu-abuan dari kanker (texture).

Sebagai saran dari penelitian berikutnya yaitu menggunakan hasil ini sebagai preproses dari algoritma penggalian data lainnya. Hal ini bertujuan untuk meningkatkan akurasi dan kecepatan pemrosesan dari diagnosis tingkat keparahan kanker payudara sehingga dapat meminimalisir waktu dan tingkat kematian yang disebabkan oleh terlambatnya diagnosis tingkat keparahan kanker payudara.

DAFTAR PUSTAKA

[1] Agresti, Alan. *Logistic Regression.* [Online] September 26, 2002. [Cited: September 2, 2014.] http://userwww.sfsu.edu/efc/classes/biol710/logistic/logisticreg.htm.

[2] ----------. 1990. *Categorical Data Analysis.* New York : John Wiley and Sons, 1990.

[3] D-13. Di Indonesia, Kasus Kanker Payudara dan Serviks Tertinggi. *Beritasatu.* [Online] Beritasatu, February 5, 2014. [Cited: September 3, 2014.] http://www.beritasatu. com/kesehatan/164592-di-indonesia-kasus-kanker-payudara-dan-serviks-tertinggi.html.

[4] Febrida, Melly. WHO: Jumlah Kematian akibat Kanker di Dunia Meningkat. *Liputan6.* [Online] Liputan6, December 16, 2013. [Cited: September 3, 2014.] http://health.liputan6.com /read/776217/who-jumlah-kematian-akibat-kanker-di-dunia-meningkat.

[5] Garson, David. Logistic Regression. [Online] 2011. [Cited: September 2, 2014.] http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm#sigtests.

[6] Kyngas H. and Rissanen M. Journal of Clinical Nursing [Journal]. - [s.l.]: Blackwell Science, Ltd., Vol. 10, 2001.

[7] Mangasarian, Olvi L and Holberg, William H. 1992. Machine Learning for Cancer Diagnosis and Prognosis. *University of Wisconsin - Madison.* [Online] 1992. [Cited: November 1, 2013.] http://pages.cs.wisc.edu/~olvi/uwmp/cancer.html.

[8] Nawawi, Qalbinur. Deteksi Dini Kanker Payudara lewat Aplikasi Ponsel. *Okehealth.* [Online] Okezone, October 22, 2013. [Cited: September 3, 2014.] http://health.okezone.com /read/2013/10/22/482/885140/hah-deteksi-dini-kanker-payudara-lewat-aplikasi-ponsel.