ANALISIS PREDIKSI CHURN MENGGUNAKAN METODE *LOGISTIC REGRESSION* DAN ALGORITMA *DECISION TREE*

Cici Olivia[1], Indwiarti[2] , Yulian Sibaroni[3]

Ilmu Komputasi
Universitas Telkom
Bandung 40257, Indonesia
[1]chiciolivia@gmail.com , [2]indwindwi@gmail.com , [3]ysibaroni@gmail.com

Abstrak

Costumer Relationship Management (CRM) merupakan sebuah strategi bisnis yang berorientasi pada pelanggan, dengan tujuan untuk memaksimalkan profit perusahaan dan kepuasan pelanggan. Salah satu aplikasi dari CRM adalah prediksi Churn. Churn mempunyai makna pelanggan memutuskan untuk keluar dari provider dan beralih ke provider lainnya atau ketidaksetiaan pelanggan. Teknik klasifikasi yang digunakan untuk prediksi Churn menggunakan metode *Logistic Regression* dan *Decision* Tree. Pada *Logistic Regression* pembentukan model berbasiskan persamaan dan kurva Logistic Regression. Sedangkan pada *Decision Tree* berbasiskan pohon keputusan. Hasil penelitian menunjukkan Decision Tree menghasilkan performansi lebih baik dibandingkan *Logistic Regression* dengan nilai akurasi 94,42% dan waktu 0,064 second. Sedangkan performansi yang dihasilkan metode *Logistic Regression* dengan akurasi sebesar 80,73% dan waktu 0,935 second. Penelitian lebih lanjut menunjukkan performansi terbaik pada metode *Decisian Tree* menggunakan satu atribut tagihan.
Kata kunci: Prediksi Churn, Klasifikasi, *Logistic Regression, Decision Tree.*

Abstract

Customer Relationship Management (CRM) is a business strategy that is oriented to the customer, with the aim of maximizing corporate profits and customer satisfaction. One of CRM application is Churn Prediction. Churn happens when custumers decided to exit from a provider and switch to another provider or disloyalty of customers. Classification techniques that used to Churn prediction are Logisitic Regression and Decision Tree methods. Building a model by Logistic Regression is based on Logistic Regression equations and curves, while on Decision Tree is based on Decision Tree. The result of Decision Tree has better performance than Logistic Regression with 94.42% of accuracy rate and 0.064 second of time. While the result of performance Logistic Regression method is 80.73% of accuracy rate and a 0,935 second of time. The further research showed the best performance of Decision Tree method uses one attribute namely bill.

Keywords*: Churn Prediction, Classification, Logistic Regression, Decision Tree.*

I. PENDAHULUAN

Besarnya persaingan antar perusahaan membuat suatu perusahaan untuk berpikir lebih maju beberapa langkah dari pesaing, agar perusahaan tersebut tetap bisa bertahan kedepannya. Salah satu perusahaan yang selalu ketat dengan persaingan adalah perusahaan dibidang telekomunikasi. Untuk meningkatkan penjualan produk suatu perusahaan, dapat dilakukan dengan mempertahankan pelanggan lama agar tidak beralih menggunakan produk perusahaan lain. Untuk itu perusahaan membutuhkan strategi khusus agar pelanggan tetap menggunakan produknya. Strategi khusus dapat dilakukan ketika perusahaan dapat memprediksi jumlah pelanggan yang akan berhenti menggunakan produk dari perusahaan tersebut (prediksi *Churn).* Penelitian tentang prediksi *Churn* sendiri sudah dilakukan sebelumnya oleh Yati Rohayati, dkk (2013) [1] dengan algoritma Decision tree dan K-Means, tetapi pada penelitian tersebut tidak melakukan perbandingan kedua metode, melainkan menyempurnakan kedua metode dengan mengetahui segmentasi pelanggan. Oleh karena itu dalam penelitian ini, dilakukan analisis prediksi *Churn* menggunakan metode *Decision Tree* dan *Logistic Regression.* Dengan data pelanggan pengguna internet yang dijadikan sebagai data training dan data testing, lalu dilakukan perbandingan kedua metode, sehingga didapatkan performansi dari kedua metode. Pengujian ketepatan prediksi menggunakan kurva ROC dan *confusion* matrix, sehingga dapat diketahui model prediksi tergolong bagus atau tidak. Tujuan

penelitian ini diharapkan dapat membantu manajemen dalam mengambil tindakan disaat adanya pelanggan yang terdeteksi *Churn*.

II. LANDASAN TEORI

A. Data Mining

Data mining adalah suatu proses ekstraksi atau penggalian data dan informasi dengan volume besar, yang belum diketahui sebelumnya, namun dapat dipahami dan berguna, serta didapatkan dari sebuah database berkapasitas besar serta digunakan untuk membuat suatu keputusan bisnis yang sangat penting[3].

B. Decision Tree

Pohon keputusan merupakan representasi sederhana dari teknik klasifikasi untuk sejumlah kelas berhingga, dimana simpul internal maupun simpul akar ditandai dengan nama atribut, rusuk-rusuknya diberi label nilai atribut yang mungkin dan simpul daun ditandai dengan kelas-kelas yang berbeda. Dalam pohon keputusan, setiap node daun diberi label kelas[3]. Manfaat utama dari penggunaan pohon keputusan adalah kemampuannya untuk mem-*break down* proses pengambilan keputusan yang kompleks menjadi lebih simpel. sehingga pengambil keputusan akan lebih menginterpretasikan solusi dari permasalahan [4].

C. Logistic Regression

Logistic Regression adalah bagian dari metode statistik yang disebut juga dengan *generalized linear models*[5]. *Logistic Regression* model digunakan saat variabel respon mengacu pada dua nilai. Misal, ketika subjek berupa benda mati atau tidak hidup, punya atau tidak memiliki sebuah karakteristik khusus dan sebagainya. Kita misalkan variabel respon sebagai y dan sebuah subjek / event y=1 ketika subjek itu memiliki karakteristik dan y=0 ketika tidak memilikinya [5]. Berikut persamaan dari *Logistic regression* [6]

$$f(z) = \frac{1}{1+e^{-z}}$$

Dimana z :

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

D. Confusion Matrix

Confusion matrix adalah sebuah matriks yang menunjukkan jumlah prediksi yang benar dan tidak benar yang dibuat oleh model klasifikasi dibandingkan dengan hasil aktual (nilai target) dalam data. Matriks nxn, dimana N adalah jumlah nilai target (kelas)[2].

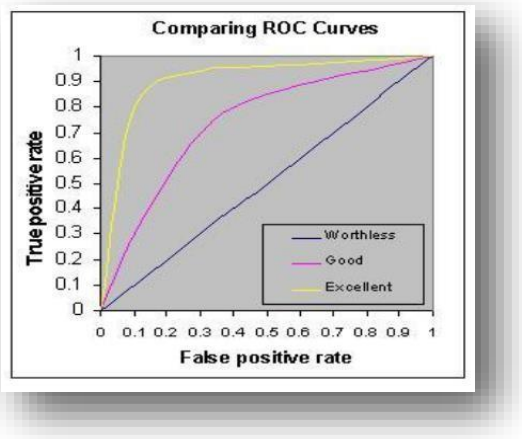
Tabel 1. Confusion Matrix

		Aktual	
		Tidak Churn	Churn
Prediksi	Tidak Churn	a	b
	Churn	c	d

TP (True Positif) = d/(c+d) x 100%
FP(False Positif) = b/(a+b) x 100%
TN (True Negatif)= a/(a+b) x 100%
FN (False Negatif)= c/(c+d) x 100%
Presisi= d/(b+d) x 100%
Akurasi = (a+d)/(a+b+c+d) x 100%

E. Kurva ROC *(Receiver Operating Characteristic)*

Penggunaan kurva ROC adalah untuk menerangkan ketetapan uji dalam berbagai tingkatan titik potong TP rate yang sesuai dengan FP rate yang ada. Kurva ROC dapat menentukan parameter model yang diinginkan sesuai dengan karakteristik dari model *classifier*. Kurva ROC menunjukan hubungan antara uji sensitifitas dan spesifisitas.[7].



Gambar 1. Kurva ROC[8]

Nilai Kurva ROC [7]:

a. 90 – 100 = Sangat Bagus (A)
b. 80 – 90 = Bagus (B)
c. 70 – 80 = Cukup(C)
d. 60 – 70 = Jelek (D)
e. 50 – 60 = Salah (E)

III. PERANCANGAN SISTEM

A. Data

Data yang digunakan pada penelitian ini sebanyak 10000 record dengan 5 atribut. Berikut informasi dari atribut data yang digunakan:

Tabel 2. Keterengan Atribut Data

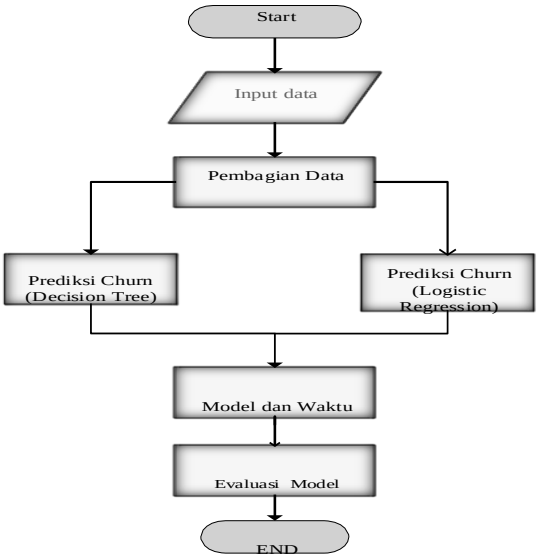
Nama Atribut	Deskripsi
NDOS	Banyaknya Produk Speedy yang digunakan
Tagihan	Tagihan Pemakaian (perbulan)
WT	Waktu Berlangganan
paket	Jenis paket yang digunakan (Mbps)
FregG	Rata-rata gangguan

B. Flowchart Sistem

Alurnya secara garis besar dapat dijelaskan sebagai berikut ini:

a. Input data, yakni melakukan pengimputan data. Data yang diinputkan adalah data yang sudah dilakukan prepocessing data.
b. Pembagian Data, yaitu melakukan pemilihan presentase antara data training dan data testing.
c. Pemilihan metode, pada tahap ini dilakukan pemilihan metode yang akan digunakan.
d. Prediksi Churn, yakni melakukan prediksi berdasarkan metode yang digunakan. Jika pada pemilihan metode yang digunakan Logisitic Regression, maka prediksi Churn yang dilakukan berdasarkan metode tersebut. Tetapi jika pemilihan metode menggunakan Decision Tree maka prediksi Churn dilakukan berdasarkan proses Decision Tree.
e. Model dan waktu, pada tahap ini akan menampilkan model dan waktu dari metode yang dipilih. Pada Decision Tree model yang didapatkan berupa pohon keputusan. Sedangkan pada Logistic Regresion model berupa persamaan dan grafik model Logistic Regression.
f. Evaluasi Model

Pada tahap ini dilakukan evaluasi menggunakan Confusion Matrix, dan kurva ROC. Evaluasi ini digunakan untuk membandingkan performansi dari Logistik Regresi dan Decision Tree.



Gambar 2. Flowchart Sistem

IV. IMPLEMENTASI SISTEM

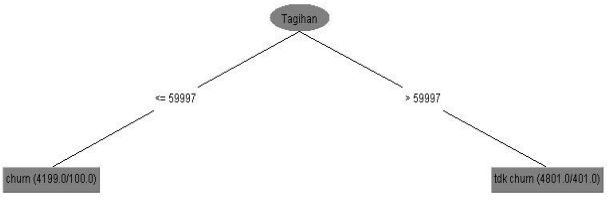
A. Percobaan

Pada implementasi sistem, dilakukan 3 kali percobaan dengan *cross validation fold* 10 untuk melihat performansi dari kedua metode, yaitu:

1. Percobaan pertama melakukan prediksi Churn menggunakan 3 atribut dengan 10 kali kemungkinan pada metode *Logisitic Regression* dan *Decision Tree.*
2. Percobaan kedua menggunakan 4 atribut dengan 5 kali kemungkinan pada *Logisitic Regression* dan *Decision Tree.*
3. Percobaan terakhir menggunakan 5 atribut dengan 1 kali kemungkinan pada *Logisitic Regression* dan *Decision Tree.*

B. Model Decision Tree

Model yang dihasilkan menggunakan 1 atribut (Tagihan) yang menghasilkan performansi terbaik. Berikut pohon keputusan yang dihasilkan:



Gambar 3. Pohon Keputusan Decision Tree

C. Evaluasi Model dan Analisis Decision Tree

Tabel 3. Percobaan Decision Tree

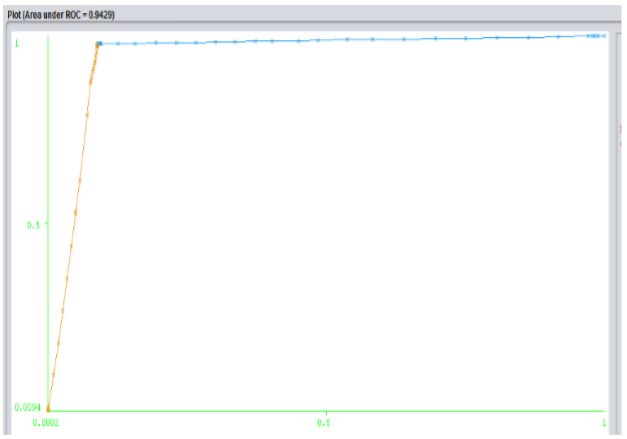
Perco -baan	Atribut	ROC (%)	Time (second)	Akurasi (%)
1	NTW	94. 1	0.423	94,42
	NTF	94. 1	0.199	94,42
	NTP	94,3	0.212	94,42
	NWF	68,5	0.184	63,46
	NWP	61	0.207	58,28
	NFP	68,6	0.184	62,95
	TWF	94, 3	0.181	94,42
	TWP	94, 4	0.197	94,42
	TFP	94, 4	0.207	94,42
	WFP	64,1	0.181	61,6
2	NTWF	94, 3	0.271	94,42
	NTWP	94, 4	0.288	94,42
	TWFP	94, 5	0.289	94,41
	TWFN	94,3	0.269	94,42
	WFPN	68,4	0.291	63,11
3	NTWFP	94,5	0.172	94.37

Berdasarkan percobaan yang dilakukan pada tabel 3 nilai akurasi tertinggi pada percobaan 1 dan 2 sebesar 94,42%, yang dihasilkan oleh kombinasi atribut NTW, NTF, NTP, TWF, TWP, TFP, NTWF, NTWP,TWFP, dan TWFN. Sedangkan akurasi dibawah 65% adalah kombinasi atribut tanpa melibatkan atribut tagihan. Sehingga pada percobaan ini atribut tagihan sangat signifikan terhadap model. Sedangkan pada percobaan ketiga nilai akurasi menjadi menurun sebesar 0,05%, artinya semakin banyak atribut yang digunakan tidak menjamin performannsi sistem yang dihasilkan lebih baik.

Dari ketiga percobaan pada tabel 4.1 performansi terbaik diperoleh pada percobaan pertama dengan kombinasi atribut tagihan, waktu berlangganan, dan ganguan (TWF) dengan akurasi sebesar 94.42 % dan waktu 0.181 second. Berikut hasil confusion matrix dan kurva ROC yang dihasilkan:

Tabel 4. Confusion Matrix Decision Tree

	Aktual	
	Tidak Churn	Churn
Prediksi	Tidak Churn	4889	111
	Churn	447	4553



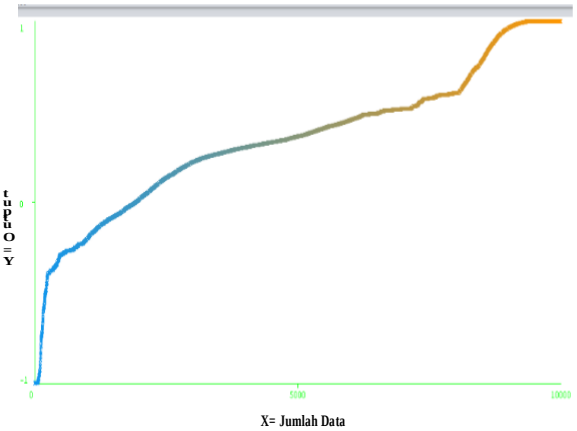
Gambar 4. Kurva ROC Decision Tree

nilai ROC yang didapatkan sebesar 94,3%, artinya model yang dihasilkan sangat bagus.

D. Model *Logistic Regression*

Model yang dihasilkan menggunakan 3 atribut yang menghasilkan performansi terbaik dengan kombinasi atribut tagihan, gangguan, dan paket (TFP). Persamaan Logisitic Regression didapatkan dari nilai koefisien Tagihan 0, FregG 0,2064, Paket -0,1175, dan intercept -2,1178. Berikut persamaan dan kurva yang dihasilkan:

$$f(z) = \frac{1}{1+e^{-(-2.1178+0.x_1+0.,02064x_2+(-0,1175x_3)}}$$



Gambar 5. Kurva Persamaan *Logistic Regression*

E. Evaluasi Model dan Analisis *Logistic Regression*

Tabel 5. Percobaan Logistic Regression

Percobaan	Atribut	ROC (%)	Time (second)	Akurasi (%)
1	NTW	94,4	1,374	79,3
	NTF	92,1	1,473	80,6
	NTP	94,3	1,139	79.6
	NWF	66,8	0.837	62,18
	NWP	58,4	0.844	54,58
	NFP	67	0.802	62,17
	TWF	92,4	1,073	80,29
	TWP	94,4	0,954	79,26
	TFP	92,1	0.935	80,73
	WFP	62,5	0,69	61,65
2	NTWF	92,1	1,241	80,57
	NTWP	94,3	1,263	79,57
	TWFP	92,1	1,265	80,69
	TWFN	92,1	1,224	80,57
	WFPN	67	0,926	62,17
3	NTWFP	91,9	2,302	81.11

Dari ketiga percobaan yang dilakukan pada tabel 5 akurasi perocobaan 1 dengan kombinasi atribut TFP tidak terlalu jauh dengan akurasi percobaan 3 (NTWFP). Sehingga dilakukan uji proporsi untuk menentukan kombinasi atribut yang digunakan. Hipotesis yang digunakan sebagai berikut:

$$H_0 : \mu_{TFP} = \mu_{NTWFP}$$
$$H_1 : \mu_{TFP} \neq \mu_{NTWFP}$$
$$p > \alpha(0{,}05) \rightarrow H_1 ditolak$$

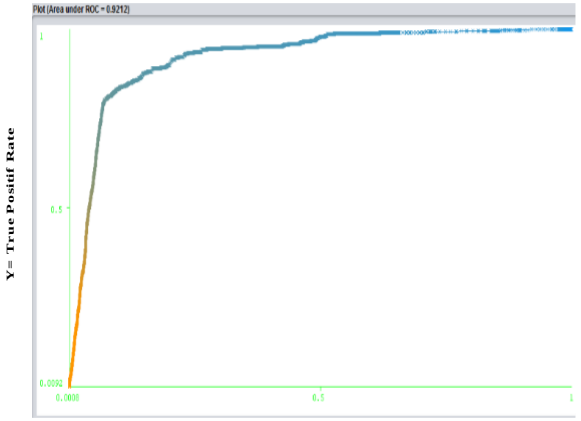
Tabel 6. Uji Proporsi Atribut

	X	N	P-Value
TFP	8070	10000	0,000
NTWFP	8110	10000	0,000

Berdasarkan uji proporsi yang dilakukan H_0 untuk atribut TFP dan NTWFP diterima. Artinya kedua kombinasi bisa digunakan, tetapi dilihat dari sisi waktu kombinasi atribut tagihan, gangguan, dan paket (TFP) menghasilkan waktu lebih kecil dibandingkan menggunkan 5 atribut. Sehingga prediksi Churn dengan performansi terbaik *Logistic Regression* menggunakan 3 atribut dengan kombinasi atribut tagihan, gangguan, dan paket (TFP). Berikut hasil *confusion matrix* dan Kurva ROC yang dihasilkan oleh akurasi tertinggi (TFP).

Tabel 7. *Confusion Matrix Logistic Regression*

		Aktual	
		Tidak Churn	Churn
Prediksi	Tidak Churn	3355	1645
	Churn	282	4718

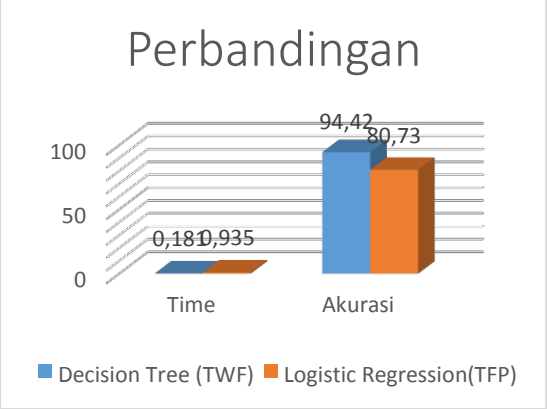


Gambar 6. Kurva ROC *Logistic Regression*

nilai ROC yang didapatkan sebesar 92,1%, model yang dihasilkan sangat bagus.

F. Perbandingan Metode Logistic Regression dan Algoritma *Decision Tree*

Berdasarkan percobaan yang telah dilakukan, selanjutnya dilakukan perbandingan performansi terbaik yang dihasilkan oleh *Logistic Regression* dan *Decision Tree*. Pada *Decision Tree* performansi terbaik diperoleh dengan kombinasi atribut tagihan, waktu berlangganan, dan ganguan (TWF) sebesar 94.42 % dengan waktu 0.181 second. Sedangkan pada *Logistic Regression* performansi terbaik diperoleh dengan kombinasi atribut tagihan, ganguan , paket (TFP) sebesar 80,73% dengan waktu 0.935. Sehingga pada prediksi Churn dengan *Decision Tree* menghasilkan performansi terbaik dibandingkan *Logisitic Regression*.



Gambar 7. Perbandingan *Decision Tree* dan *Logistic Regression*

Melihat atribut tagihan sangan signifikan maka dilakukan percobaan lebih lanjut menggunakan dua atribut dimana satu atribut yang digunakan adalah atribut tagihan untuk melihat atribut lainnya yang berpengaruh terhadap akurasi dan waktu yang dihasilkan.

Tabel 8. Kombinasi Dua Atribut

Metode	Atribut	Time (second)	Akurasi (%)
Decision Tree	TN	0,304	94,42
	TF	0,127	94,42
	TW	0,107	94,42
	TP	0.837	94,42
	T	0,064	94,42

Dari semua kombinasi dua atribut nilai akurasi yang dihasilkan metode Decision Tree tidak terjadi perubahan, begitu juga dengan menggunakan satu atribut tagihan. Artinya atribut lain tidak berpengaruh dalam prediksi Churn pada data ini.

Tabel Confusion Matrix 1 Atribut:

	Aktual		
	Tidak Churn		Churn
Prediksi	Tidak Churn	4889	111
	Churn	447	4553

V. KESIMPULAN

Berdasarkan analisis terhadap pengujian yang dilakukan pada sistem prediksi Churn menggunakan metode Logistic Regression dan Algoritma Decision Tree, maka dapat ditarik kesimpulan sebagai berikut:

1. Performansi terbaik dalam prediksi Churn menggunakan metode Decision Tree dengan 1 atribut menghasilkan akurasi sebesar 94,42 % dan waktu 0.935 second.
2. Atribut tagihan sangat signifikan terhadap model prediksi Churn.
3. Pada Logistic Regression performansi terbaik diperoleh dengan menggunakan 3 atribut dengan kombinasi tagihan, rata-rata gangguan perbulan, dan paket (TFP) sebesar 80,73% dengan waktu 0.935.
4. Penambahan atribut tidak menjamin performansi yang dihasilkan lebih baik.

REFERENSI

[1] Syamala, M. P., & Rohayati, Y. (2013). Analisis Prediksi Churn Dan Segmentasi Pelanggan Speedy Retail Daerah Operasional Bandung Menggunakan Algoritma Decision Tree Dan K-Means.

[2] Witten, Ian H.(2011). Data Mining Pratical Machine Learning Tools and Tecniques.

[3] Hermawati, F. A. (2013). *Data Mining.* (P. Christian, Penyunt.) Surabaya.

[4] Marwana. (2013). Algoritma C4.5 Untuk Simulasi Prediksi Kemenangan Dalam Pertandingan Sepak Bola. 2. Diambil kembali dari http://jim.stimednp.ac.id/wp-content/uploads/2014/03/ Algoritma C4.5 Untuk Simulasi Prediksi Kemenangan Dalam Pertandingan Sepak Bola.pdf

[5] Santoso, B. (2007). *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis.* Yogyakarta.

[6] Logistic Regression, Wikipedia, diambil tanggal 3 feberuari 2015 dari http://en.wikipedia.org/wiki/Logistic_regression.

[7] ROC Curve Analysis in MedCalc. (t.thn.). Diambil kembali dari w.medcalc.org: w.medcalc.org/manual/roc-curves.php

[8] The Area Under an ROC Curve. (t.thn.). Dipetik Januari 13, 2015, dari http://gim.unmc.edu: http://gim.unmc.edu/dxtests/roc3.html