



Penerapan Algoritma Klasifikasi C4.5 Berbasis AdaBoost Untuk Prediksi Loyalitas Pelanggan Distributor Pulsa Elektronik

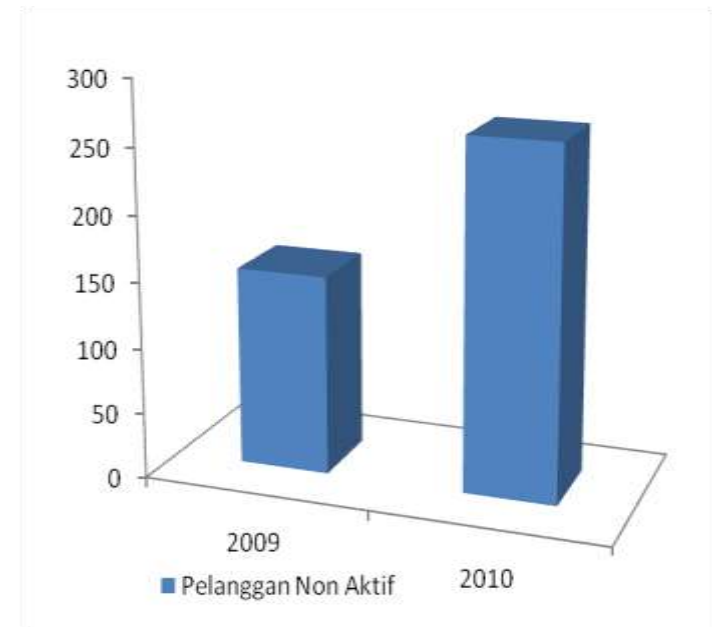
Eni Irfiani
14000278

Latar Belakang Masalah

- Distributor adalah badan perseorangan yang membeli produk dalam jumlah besar dan menyalurkan kepada **sub distributor** atau kepada **pengecer langsung**
- **Distributor pulsa elektronik menjual pulsa isi ulang elektronik** prabayar GSM/CDMA yang banyak diminati oleh pelanggan
- **Banyaknya pelanggan yang tidak loyal** pada pelanggan telepon selular pasca bayar (Govindaraju, dkk, 2008)
- **Meningkatnya jumlah pelanggan yang tidak aktif** atau berpindah ke kompetitor lain (Liao, 2007)
- **Akurasi prediksi** loyalitas pelanggan pada perusahaan telekomunikasi selular pascabayar **71.31% → fair classification** (Govindaraju, 2008) & (Goronescu, 2011)

Latar Belakang Masalah

Tahun	Jumlah Pelanggan	Pelanggan Non Aktif
2009	535	28%
2010	780	33%



Grafik peningkatan pelanggan non aktif
(sumber: distributor pulsa elektronik Garuda Reload Depok)

Latar Belakang Masalah

- Perlu diketahui dan diprediksi pola pelanggan loyal dan tidak dari aktifitas penjualan pulsa elektronik
- Pencarian dan prediksi pola dari data menggunakan metode klasifikasi data mining
- Penelitian yang telah dilakukan dengan menggunakan c4.5 untuk prediksi loyalitas pelanggan → belum akurat

Rumusan Masalah

Sebagian besar penerapan algoritma C4.5 untuk prediksi loyalitas pelanggan, tingkat akurasi masih belum di level excellence

Pertanyaan penelitian:

Seberapa meningkat akurasi Algoritma C4.5 berbasis AdaBoost dapat memprediksi loyalitas pelanggan?



Tujuan Penelitian

Menerapkan metode adaboost pada Algoritma C4.5 untuk meningkatkan akurasi prediksi loyalitas pelanggan



Manfaat Penelitian

- Mempersiapkan treatment khusus untuk pelanggan yang tidak loyal, sehingga mereka jadi loyal
- Meningkatkan pendapatan perusahaan karena treatment khusus akan meningkatkan loyalitas



Ruang Lingkup

- **Research:** Penelitian dibatasi pada pengujian model algoritma klasifikasi c4.5 dan metode boosting Adaboost
- **Data:** Distributor Garuda Reload Depok dimana observasi dilakukan selama tiga bulan (April-Juni 2011)

Tinjauan Studi

Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Service (Kraljevic & Gotovac, 2010)

- Penelitian untuk memprediksi pelanggan yang tidak loyal pada perusahaan telekomunikasi prabayar HT Mostar di Bosnia & Herzegovina
- 2/3 data training, 1/3 data testing
- Komparasi model Decision Tree dan Neural Network
- Model pohon keputusan (Decision Tree) memiliki akurasi paling baik

Tinjauan Studi

Data Mining in Churn Analysis Model for Telecommunication Industry (Oseman,dkk, 2010)

- Penelitian untuk mencegah berkurangnya pendapatan perusahaan yang disebabkan pelanggan yang tidak loyal
- Menggunakan model pohon keputusan algoritma ID3
- Variabel prediksi berdasarkan data demografi pelanggan serta aktivitas pelanggan menggunakan pelayanan

Tinjauan Studi

The Comparative Analysis and Study of Mobile-based Customer Data Churn Prediction Model (Jin-Hui & Jian-Jun, 2009)

- Penelitian untuk memprediksi pelanggan yang tidak loyal pada perusahaan telekomunikasi
- 100.000 data training, 100.000 data testing
- Komparasi model Decision Tree, Neural Network, SVM
- Model pohon keputusan memiliki akurasi paling baik

Tinjauan Studi

Perancangan Sistem Prediksi *Churn* Pelanggan PT. Telekomunikasi Seluler Dengan Memanfaatkan Proses Data Mining(Govindaraju, dkk, 2008)

- Penelitian untuk memprediksi pelanggan yang tidak loyal pada perusahaan telekomunikasi selular pasca bayar di Bandung
- Menggunakan model Decision Tree
- Variabel atribut yang digunakan data informasi pelanggan, data aktivitas transaksi, data pembayaran
- Akurasi prediksi 71.31%

Tinjauan Studi

Metode Pohon Gabungan: Solusi Pilihan Untuk Mengatasi Kelemahan Pohon Regresi dan Klasifikasi Tunggal (Sartono & Syafitri, 2010)

- Menggunakan metode pohon regresi dan pohon klasifikasi
- Komparasi metode untuk meningkatkan akurasi antara lain Boosting, Bagging, Random Forest
- Metode Boosting mempunyai akurasi klasifikasi terbaik



Tinjauan Pustaka

1. Loyalitas Pelanggan
2. Distributor Pulsa Elektronik
3. Algoritma C4.5
4. C4.5 berbasis AdaBoost



Loyalitas Pelanggan

Faktor pengukuran loyalitas pelanggan (Griffin, 2005) :

- Pelanggan berulang
- Melakukan pembelian kedua dalam periode waktu tertentu
- Frekuensi pembelian
- Jumlah pembelian



Distributor Pulsa Elektronik

- Distributor: membeli produk dalam jumlah besar dan menyalurkan kepada pengecer
- Pulsa Elektronik: voucher elektronik isi ulang GSM atau CDMA
- PT. Garuda Reload Depok menggunakan *multi chip server*



Algoritma C4.5

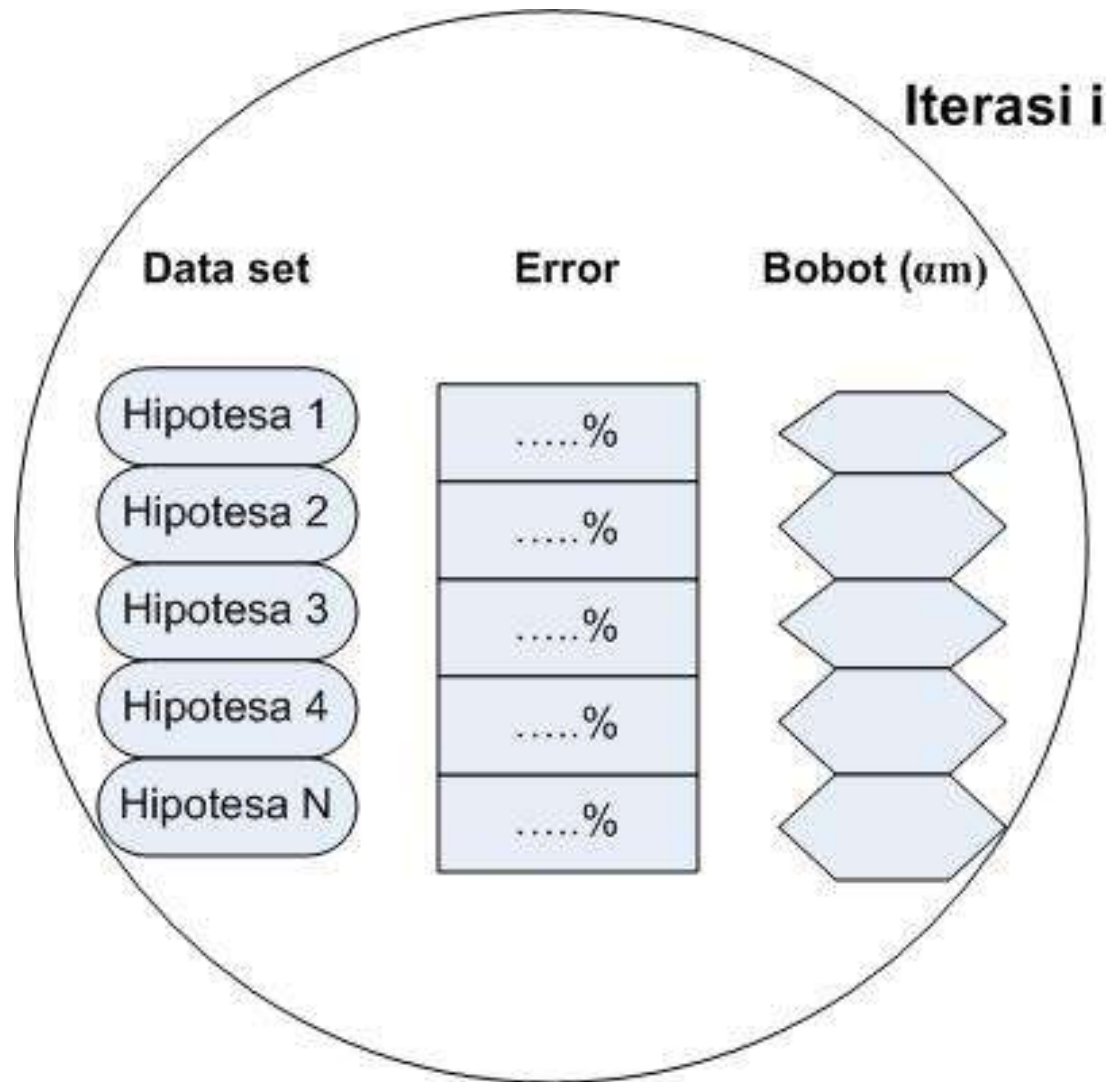
- Model Data Mining yang berbasis supervised learning
- Digambarkan dalam bentuk pohon yang bercabang untuk menyelesaikan permasalahan klasifikasi
- Digunakan untuk kasus-kasus dimana output nya bernilai diskrit



C4.5 berbasis AdaBoosting

- Metode dalam klasifikasi dan prediksi untuk meningkatkan ketelitian
- Kombinasi dari suatu model
- Pemberian bobot pada pohon tunggal

Algoritma AdaBoost



Hipotesa baru = hipotesa lama + (bobot* h)

Kerangka Pemikiran

PROBLEMS

Prediksi loyalitas pelanggan menggunakan algoritma C4.5 belum akurat



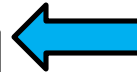
APPROACH

Algoritma C4.5 berbasis AdaBoost



DEVELOPMENT

Framework RapidMiner versi 5.1



IMPLEMENTATION

Objek: Distributor pulsa elektronik



EVALUATION

Confusion Matrix

Kurva ROC



RESULT

Algoritma C4.5 berbasis AdaBoost meningkatkan akurasi dalam memprediksi loyalitas pelanggan

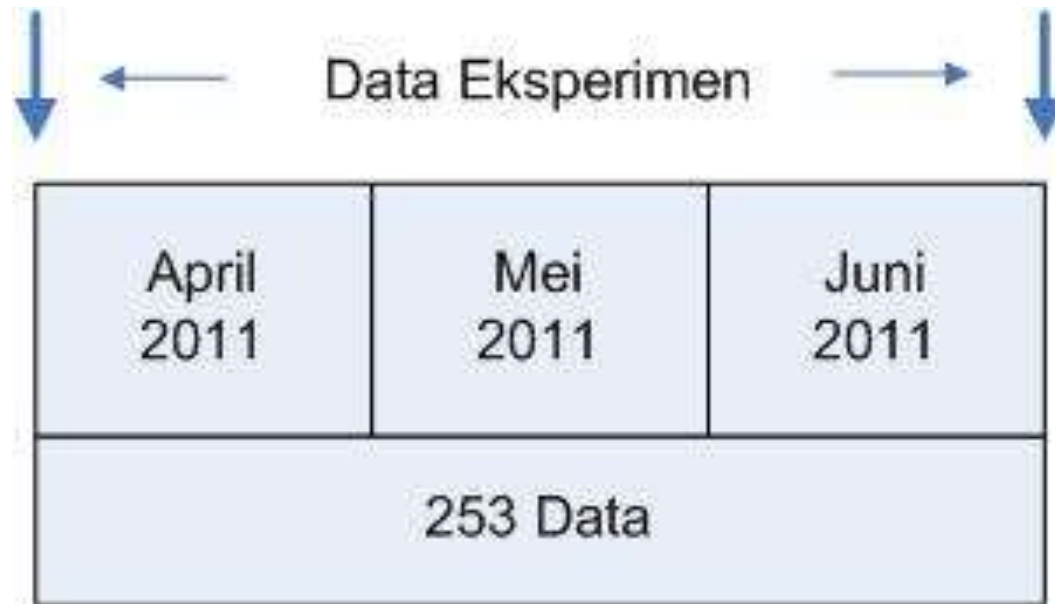


Metode Penelitian

1. Pengumpulan Data
2. Pengembangan Model
 - a. Algoritma C4.5
 - b. C4.5 berbasis AdaBoost
3. Pengujian Model pada Data Eksperimen
4. Evaluasi dan Validasi Data Eksperimen
 - a. Confusion matrix
 - b. Kurva ROC

1. Pengumpulan Data

Ekperimen menggunakan 253 diambil dari bulan april sampai juni 2011



1. Pengumpulan Data

Parameter atribut untuk memprediksi loyalitas pelanggan antara lain

Deposit	Minim
	Cukup
	Banyak
Frekuensi	Jarang
	Sedang
	Sering
Area	Desa
	Kota
Lama Langganan	PinggirKota
	Baru
	Lama

2. Pengembangan Model

Algoritma C4.5

1. Menentukan atribut , diambil dari data historis.
2. Menentukan atribut yang terpilih lebih dahulu dengan menghitung nilai entropi

$$Entropi(y) = \sum_{i=1}^n - p_i * \log_2 p_i$$

3. Menghitung nilai information gain terbesar
4. Menghitung information gain dari output y data yang dikelompokkan berdasarkan atribut A:

5. Hitung Gain (y,A)

$$Gain(y,A) = Entropi(y) - \frac{|\underline{y_c}|}{|y|} entropi(y_c)$$

6. Ulangi langkah ke-2 hingga semua atribut terpecah

2. Pengembangan Model

Algoritma C4.5 Berbasis AdaBoost

1. Menentukan awal bobot $W_i = 1/n$ (n =pengamatan)
2. Menentukan iterasi, kemudian lakukan proses:
 - a. Susun pohon tunggal dengan memperhatikan bobot W_i
 - b. Hitung tingkat kesalahan klasifikasi
 - c. Hitung $\alpha_m = \frac{1 - e_m}{e_m}$
 - d. Tentukan bobot baru untuk setiap pengamatan
3. Menentukan nilai akhir class
Dugaan kelas akhir menggunakan fungsi
 $\text{sign}(\sum \alpha_m)$
hasil dugaan kelas akhir dinotasikan +1 dan -1

3. Pengujian Model

a. Algoritma C4.5

Menentukan entropi & Gain Information dari masing-masing atribut.

Atribut	Entropy	Information Gain
Area	0.8389722	0.013432982
Deposit	0.63528352	0.21712166
Lama Berlangganan	0.83142275	0.020982427
Frekuensi	0.68080834	0.171596836

Perhitungan Node 1

Node			Jumlah kasus (S)	Aktif (S1)	Tidak (S2)	Entropy	Gain	Ket
1.1	Deposit – Cukup		18	13	5	0.85241		
	Frekuensi	Jarang	8	4	4	1	0.191	Gain Tertinggi Node 1.1
		Sedang	6	5	1	0.65002		
		Sering	4	4	0	0		
	Area	Desa	6	4	2	0.9183	0.024	
		PinggirKota	6	5	1	0.65002		
		Kota	6	4	2	0.9183		
	Lama Berlangganan	Baru	9	8	1	0.50326	0.105	
		Lama	9	5	4	0.99108		
1.2	Deposit – Minim		19	9	10	0.998		
	Frekuensi	Jarang	8	2	6	0.81128	0.293	Gain Tertinggi Node 1.2
		Sedang	7	3	4	0.98523		
		Sering	4	4	0	0		
	Area	Desa	5	2	3	0.97095	0.051	
		PinggirKota	8	5	3	0.95443		
		Kota	6	2	4	0.9183		
	Lama Berlangganan	Baru	11	6	5	0.99403	0.021	
		Lama	8	3	5	0.95443		
1.3	Deposit – Banyak		17	17	0	0	0	Aktif

Perhitungan Node 1.1

Node			Jumlah kasus (S)	Aktif (S1)	Tidak (S2)	Entropy	Gain	Ket
1.1	Deposit - Cukup							
1.1.1	Frekuensi	Jarang	8	4	4	1		
	Area	Desa	3	1	2	0.9183	0.061	
		PinggirKota	3	2	1	0.9183		
		Kota	2	1	1	1		
	Lama Berlangganan	Baru	4	3	1	0.81128	0.189	Gain Tertinggi Node 1.1.1
		Lama	4	1	3	0.81128		
1.1.2	Frekuensi	Sedang	6	5	1	0.65002		
	Area	Desa	2	2	0	0	0.317	Aktif
		PinggirKota	2	2	0	0		
		Kota	2	1	1	1		
	Lama Berlangganan	Baru	3	3	0	0	0.191	
		Lama	3	2	1	0.9183		
1.1.3	Frekuensi	Sering	4	4	0	0	0	Aktif

Perhitungan Node 1.2

Node			Jumlah kasus (S)	Aktif (S1)	Tidak (S2)	Entropy	Gain	Ket
1.2	Deposit - Minim							
1.2.1	Frekuensi	Jarang	8	2	6	0.81128		
	Area	Desa	3	1	2	0.9183	0.123	Tidak
		PinggirKota	3	1	2	0.9183		
		Kota	2	0	2	0		
	Lama Berlangganan	Baru	5	2	3	0.97095	0.204	
		Lama	3	0	3	0		
1.2.2	Frekuensi	Sedang	7	3	4	0.98523		
	Area	Desa	2	1	1	1	0.306	Gain Tertinggi Node 1.2.2
		PinggirKota	3	2	1	0.9183		
		Kota	2	0	2	0		
	Lama Berlangganan	Baru	4	1	3	0.81128	0.128	
		Lama	3	1	2	0.9183		
1.2.3	Frekuensi	Sering	4	4	0	0		Aktif

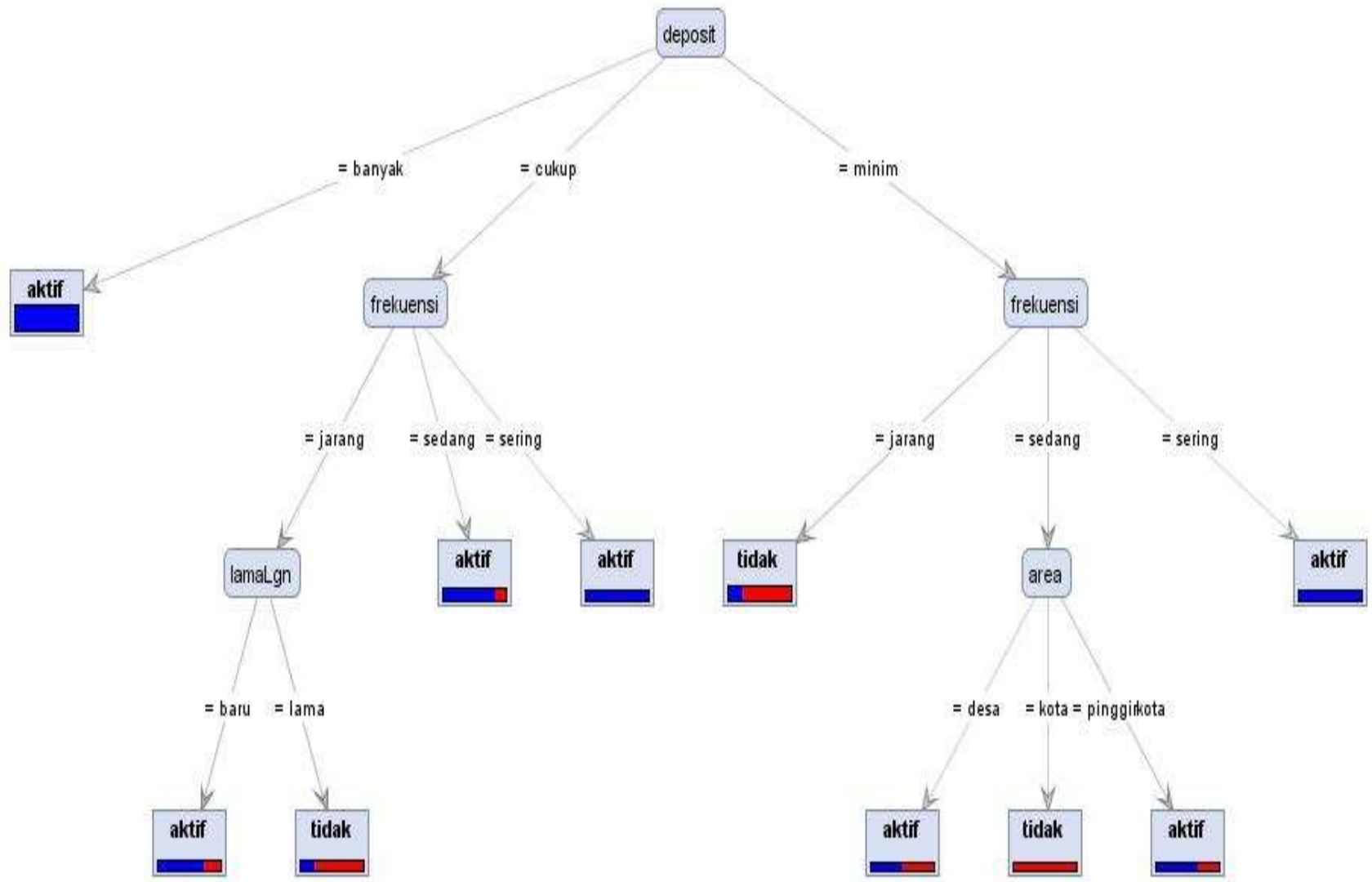
Perhitungan Node1.1.1

Node			Jumlah kasus (S)	Aktif (S1)	Tidak (S2)	Entropy	Gain	Ket
1.1	Deposit - Cukup							
1.1.1	Frekuensi - Jarang							
1.1.1.1	Lama Berlangganan	Baru	4	3	1	0.81128		
	Area	Desa	2	1	1	1	0.3113	Aktif
		PinggirKota	1	1	0	0		
		Kota	1	1	0	0		
1.1.1.2	Lama Berlangganan	Lama	4	1	3	0.81128		
	Area	Desa	1	0	1	0	0.8113	Tidak
		PinggirKota	2	1	1	1		
		Kota	1	0	1	0		

Perhitungan node 1.2.2

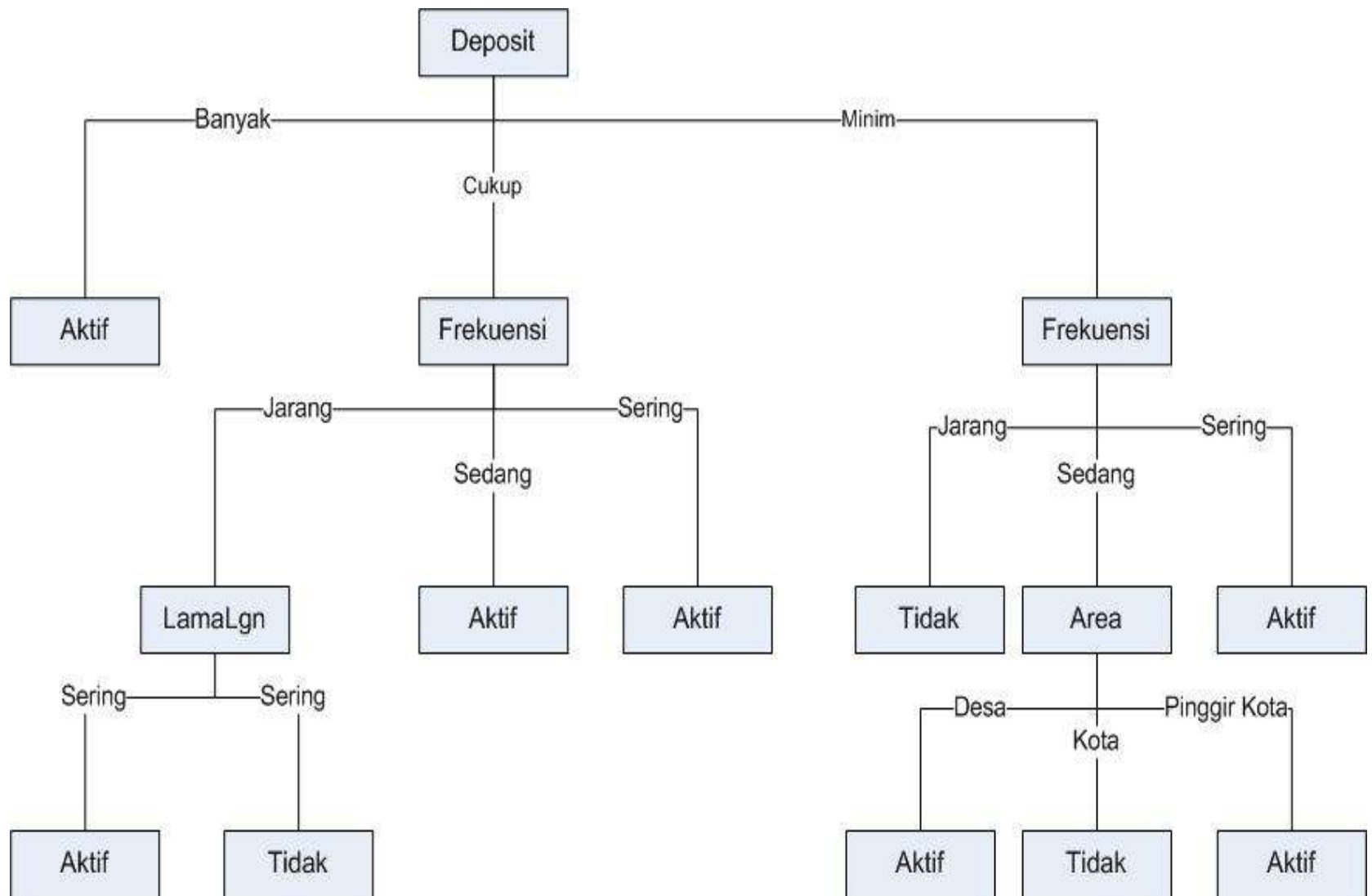
Node			Jumlah kasus (S)	Aktif (S1)	Tidak (S2)	Entropy	Gain	Ket
1.2	Deposit - Minim							
1.2.2	Frekuensi - Sedang							
1.2.2.1	Area	Desa	2	1	1	1		
	Lama Berlangganan	Baru	1	1	0	0	0	Aktif
		Lama	1	0	1	0		
1.2.2.2	Area	Pinggir Kota	3	2	1	0.9183		
	Lama Berlangganan	Baru	2	1	1	1	0.252	Aktif
		Lama	1	1	0	0		
1.2.2.3	Area	Kota	2	0	2	0		
	Lama Berlangganan	Baru	1	0	1	0	0	Tidak
		Lama	1	0	1	0		

Model pohon keputusan algoritma C4.5

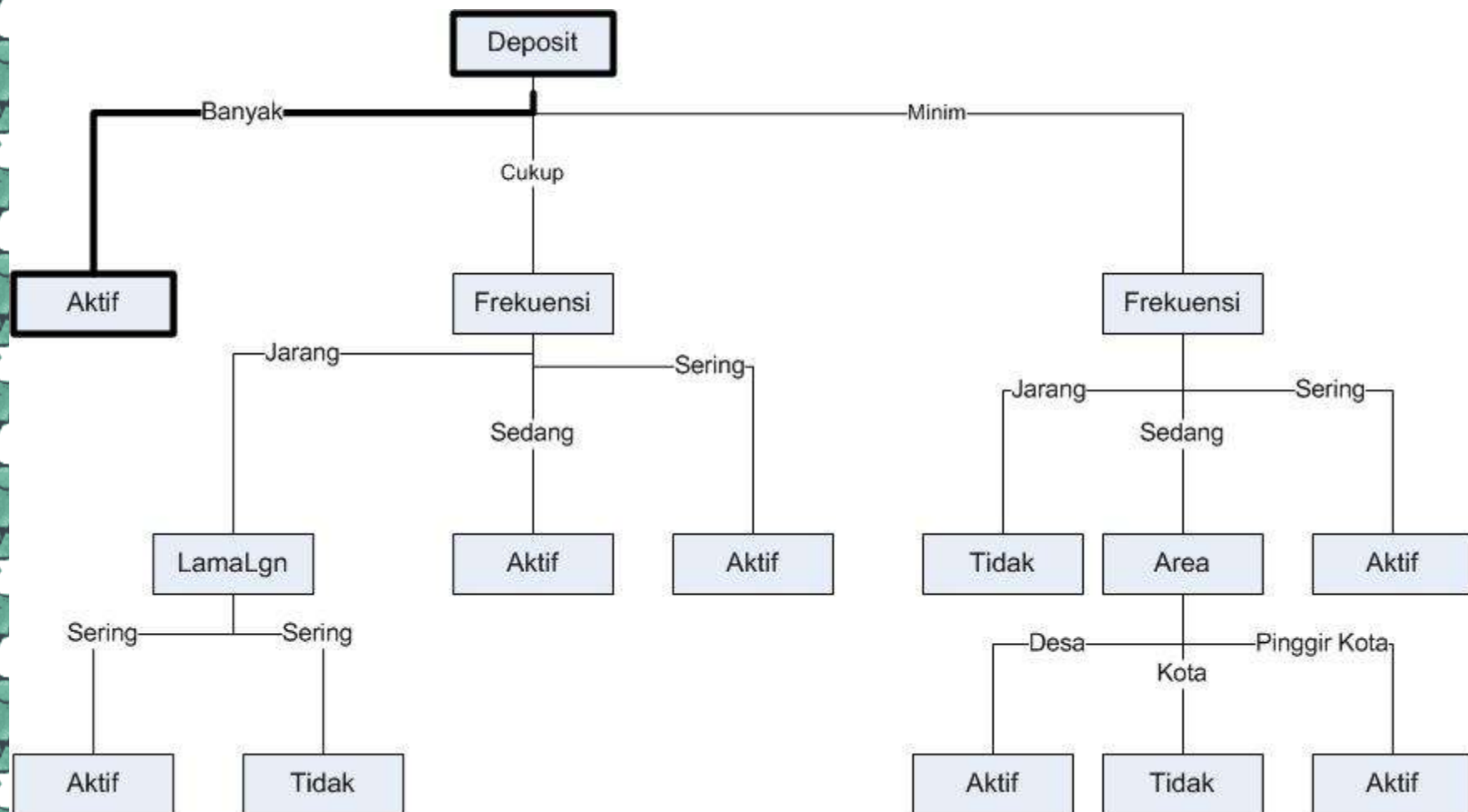


b. Algoritma C4.5 berbasis AdaBoost

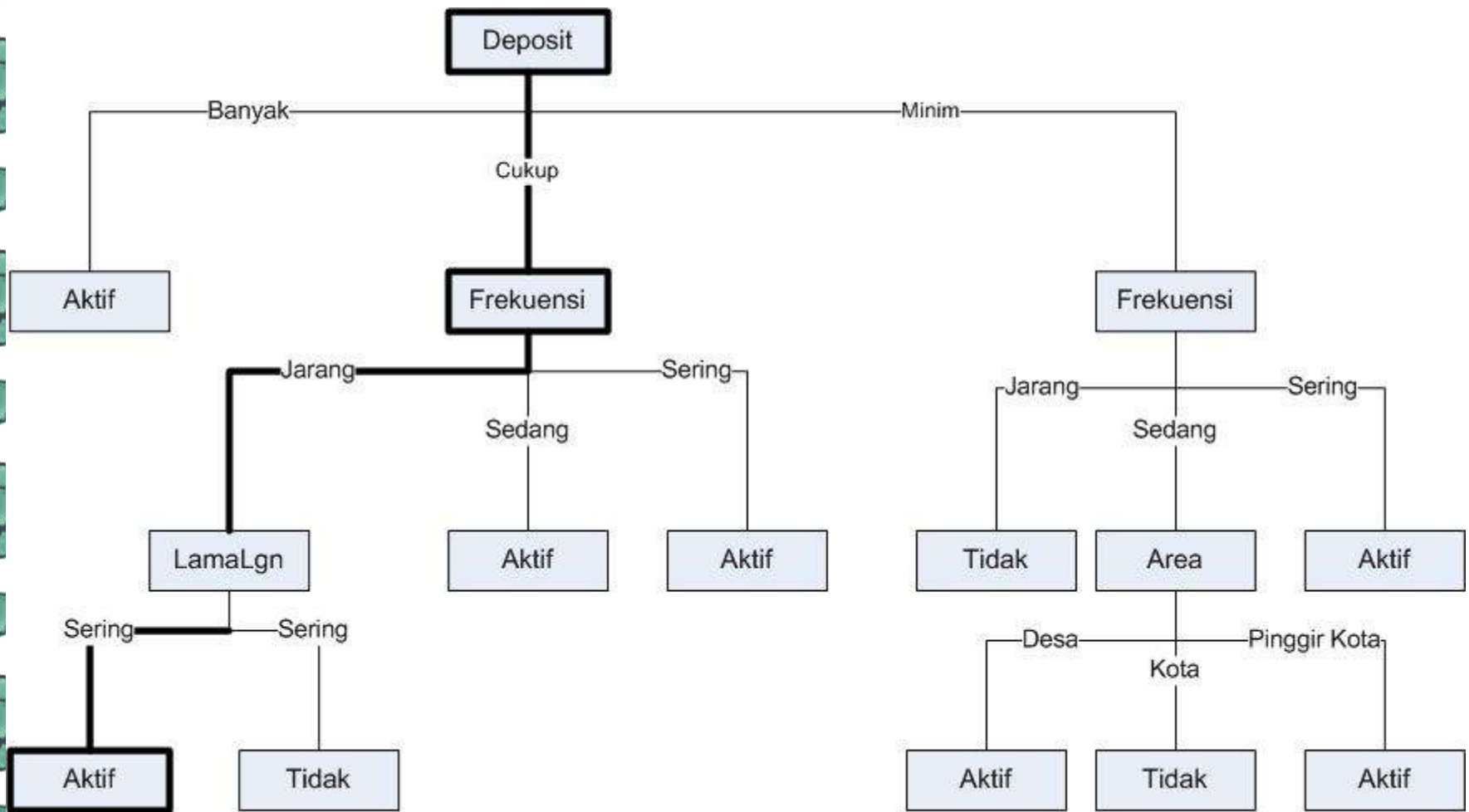
1. Menentukan bobot awal $W_i = 1/n = 1/59 = 0.0169$
2. Penentuan iterasi dimana nomor iterasi $m = 1, 2, \dots, M$. kemudian susun pohon tunggal dengan memperhatikan bobot W_i
 - a. tingkat kesalahan klasifikasi
 - b. Hitung α_m
 - c. Tentukan bobot baru untuk setiap pengamatan



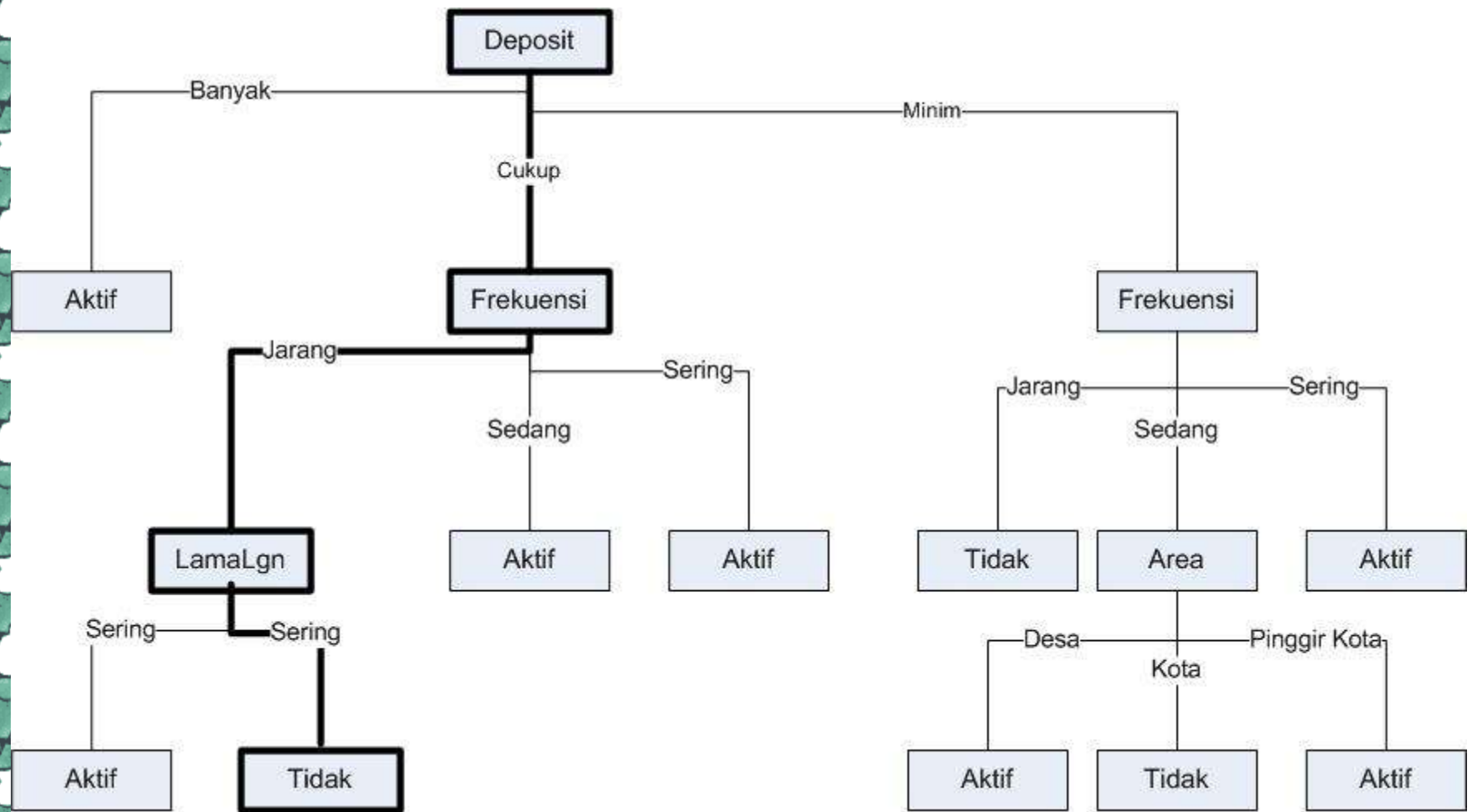
Model Pohon Keputusan Secara Lengkap



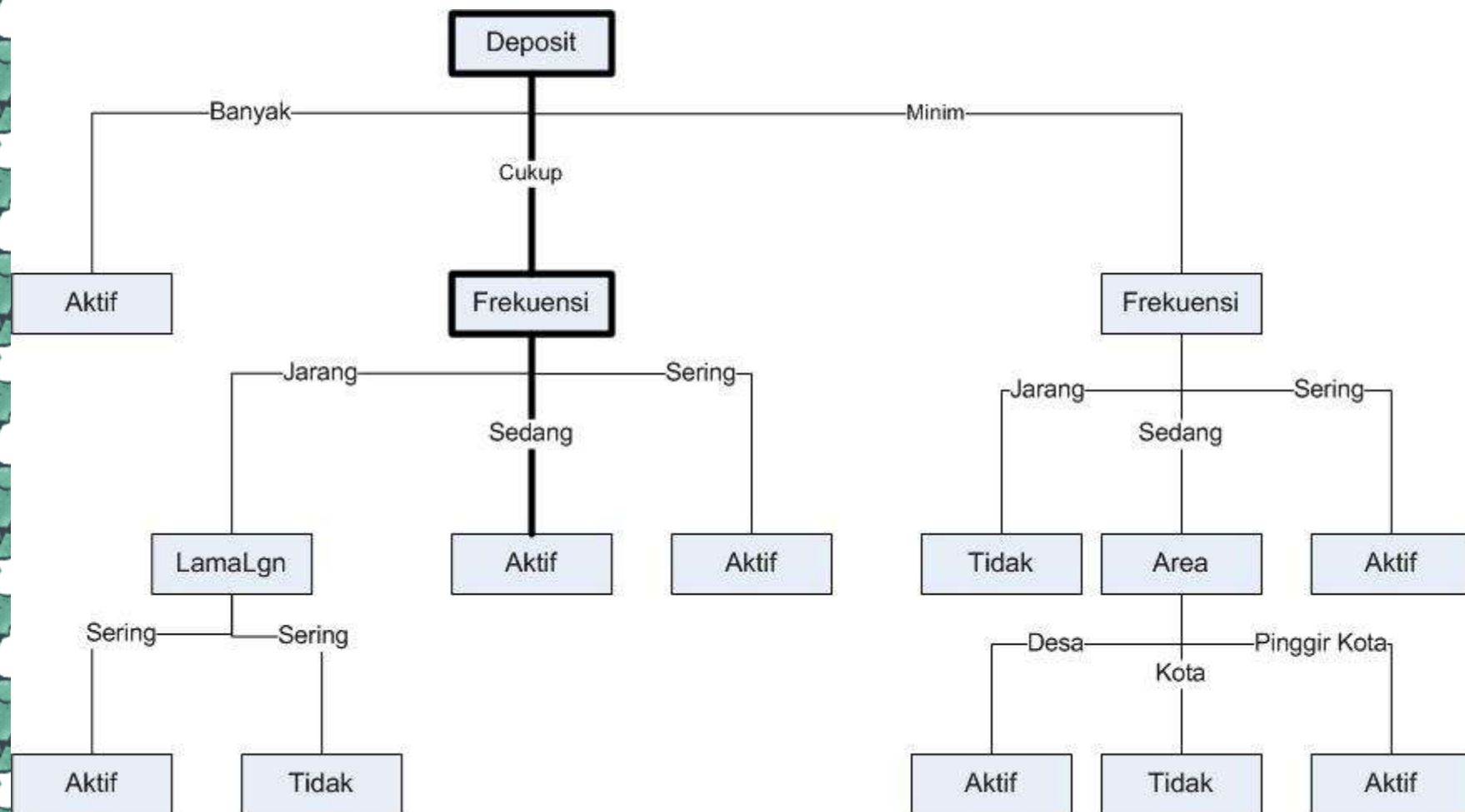
Model Pohon Keputusan hipotesa 1



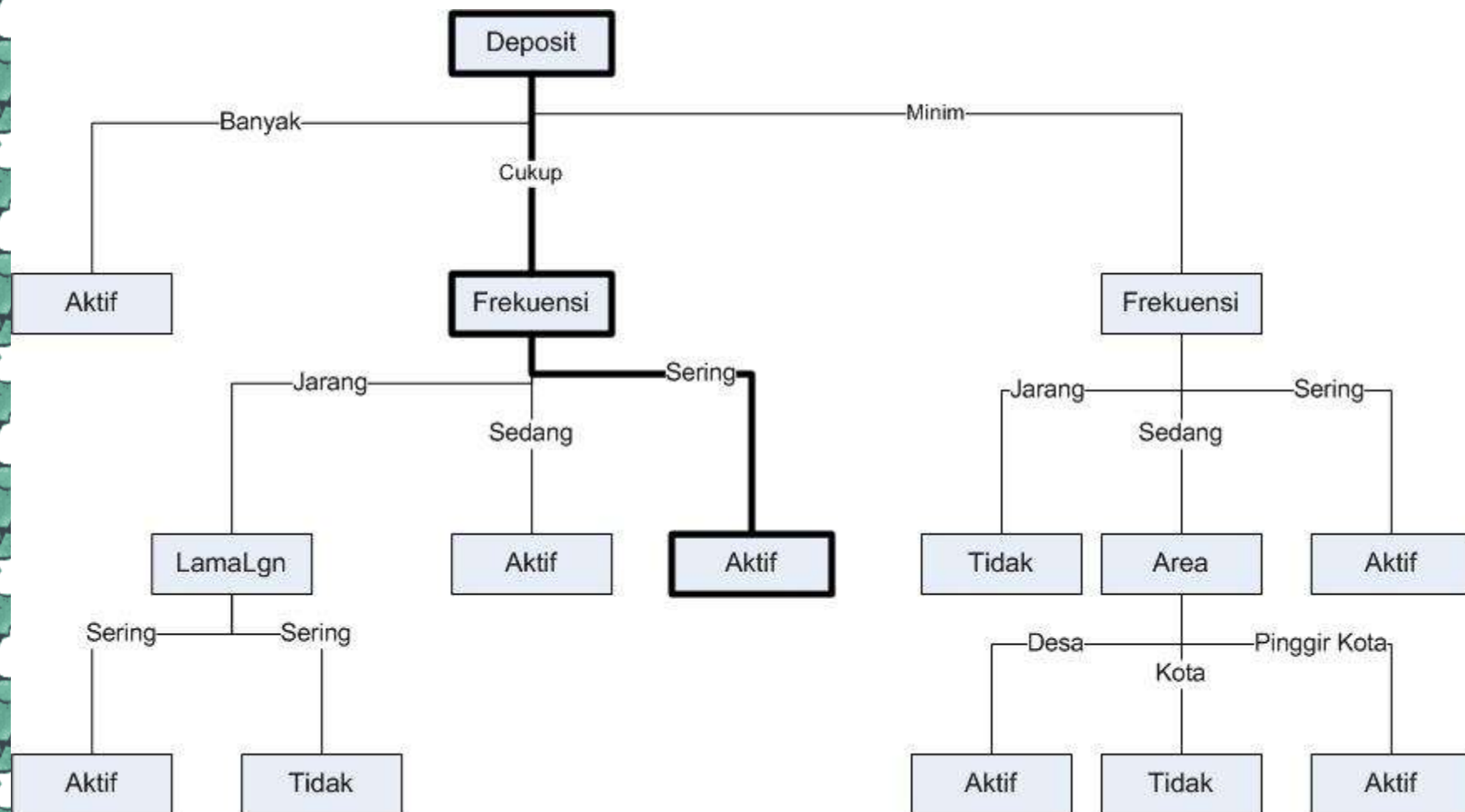
Model Pohon Keputusan hipotesa 2



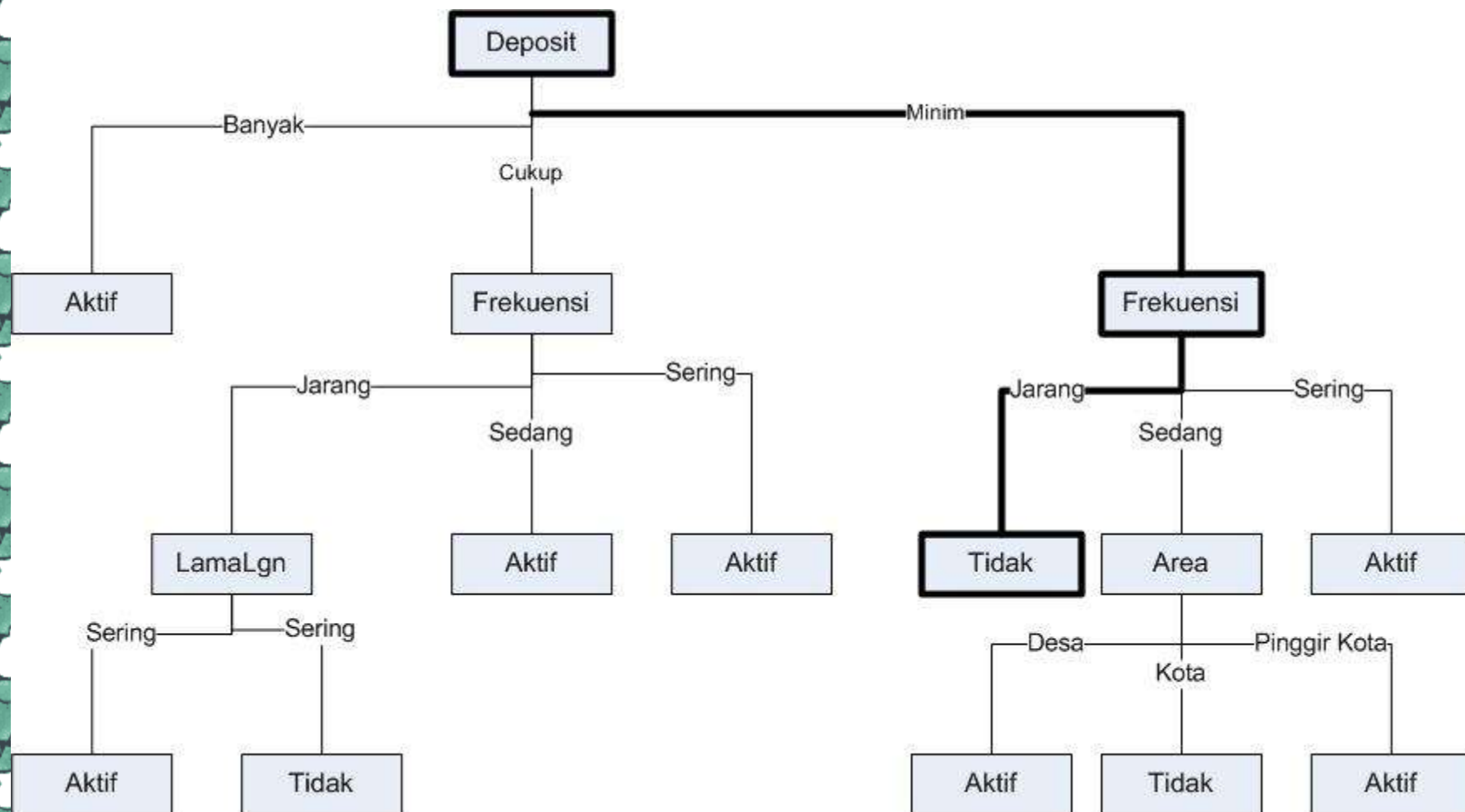
Model Pohon Keputusan hipotesa 3



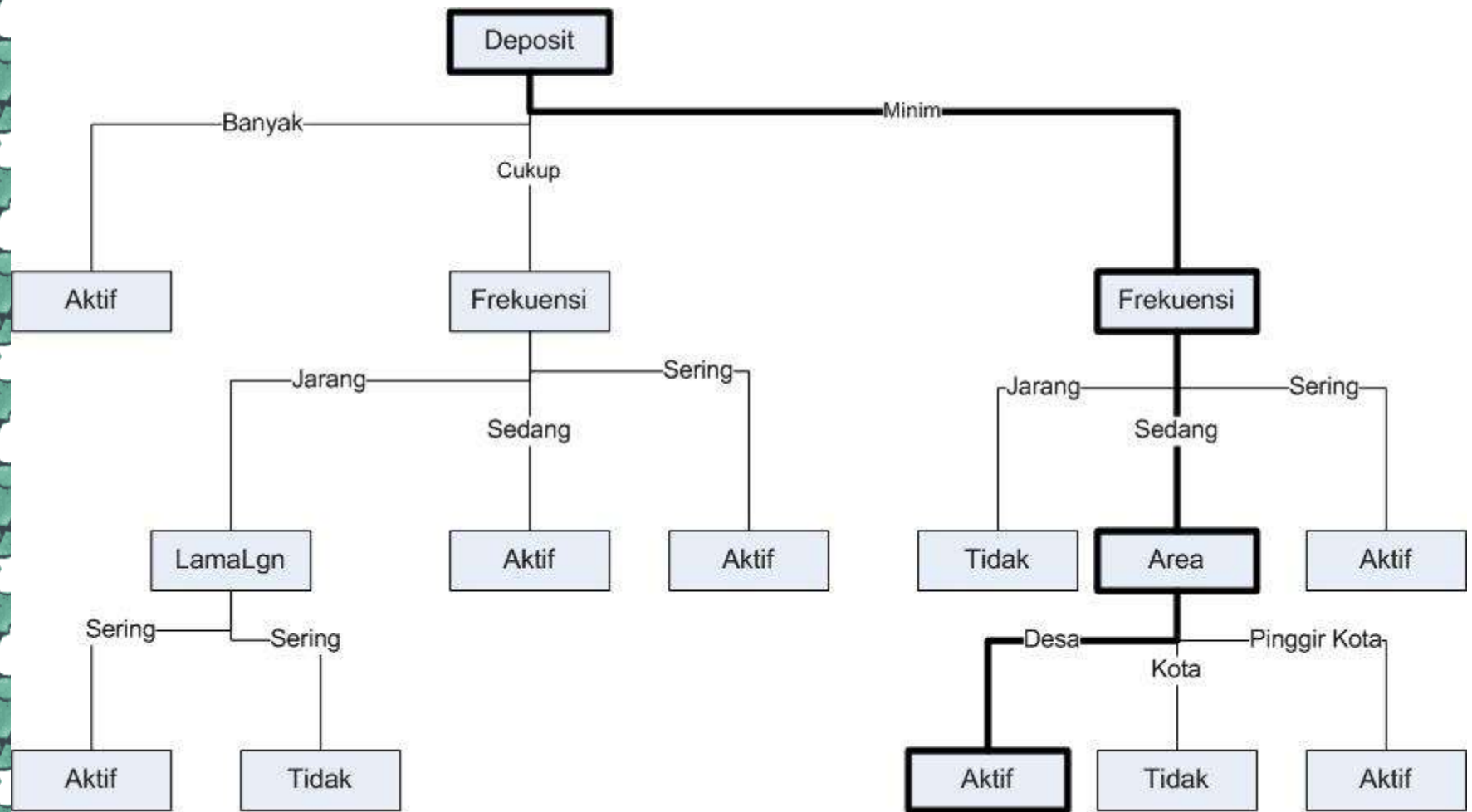
Model Pohon Keputusan hipotesa 4



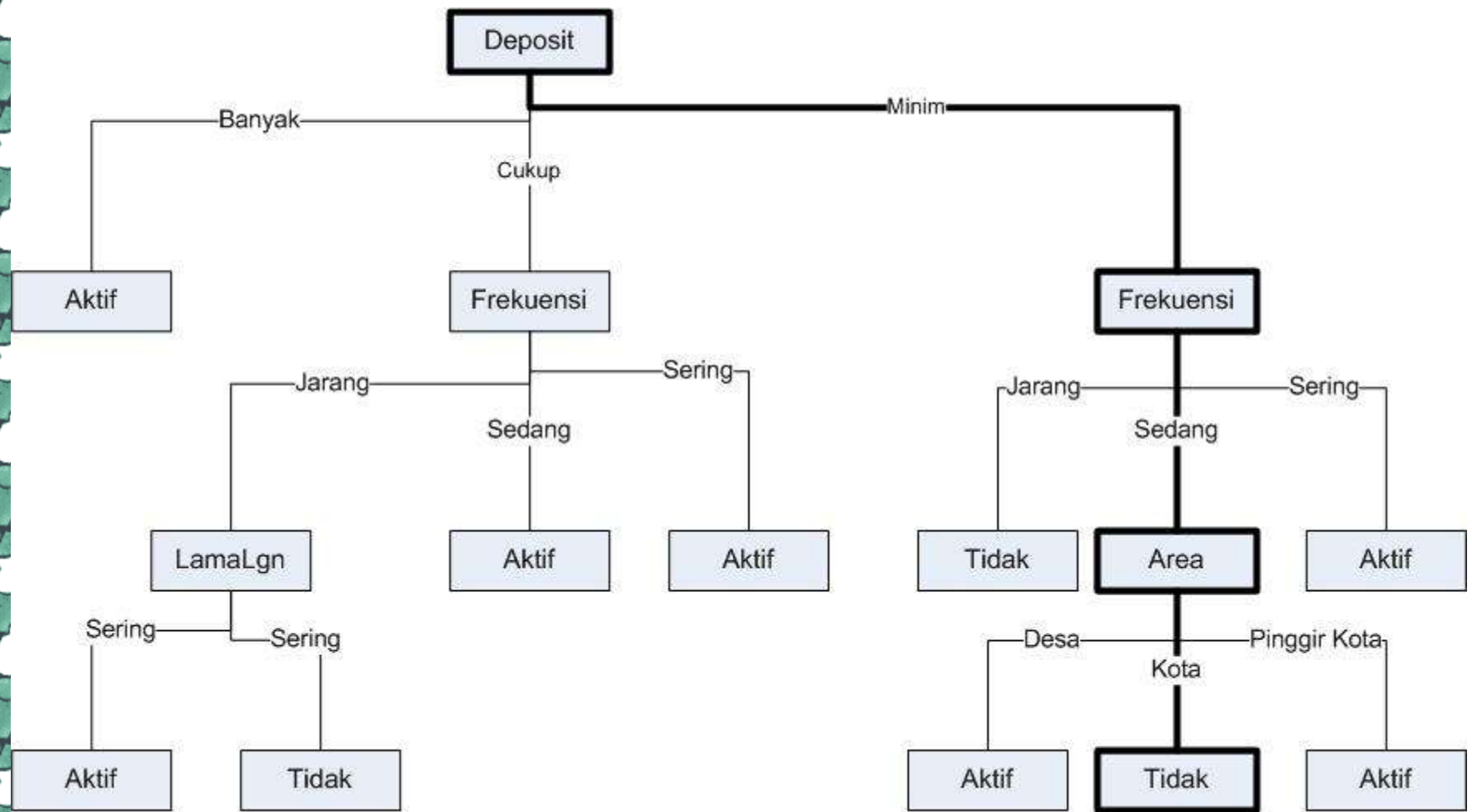
Model Pohon Keputusan hipotesa 5



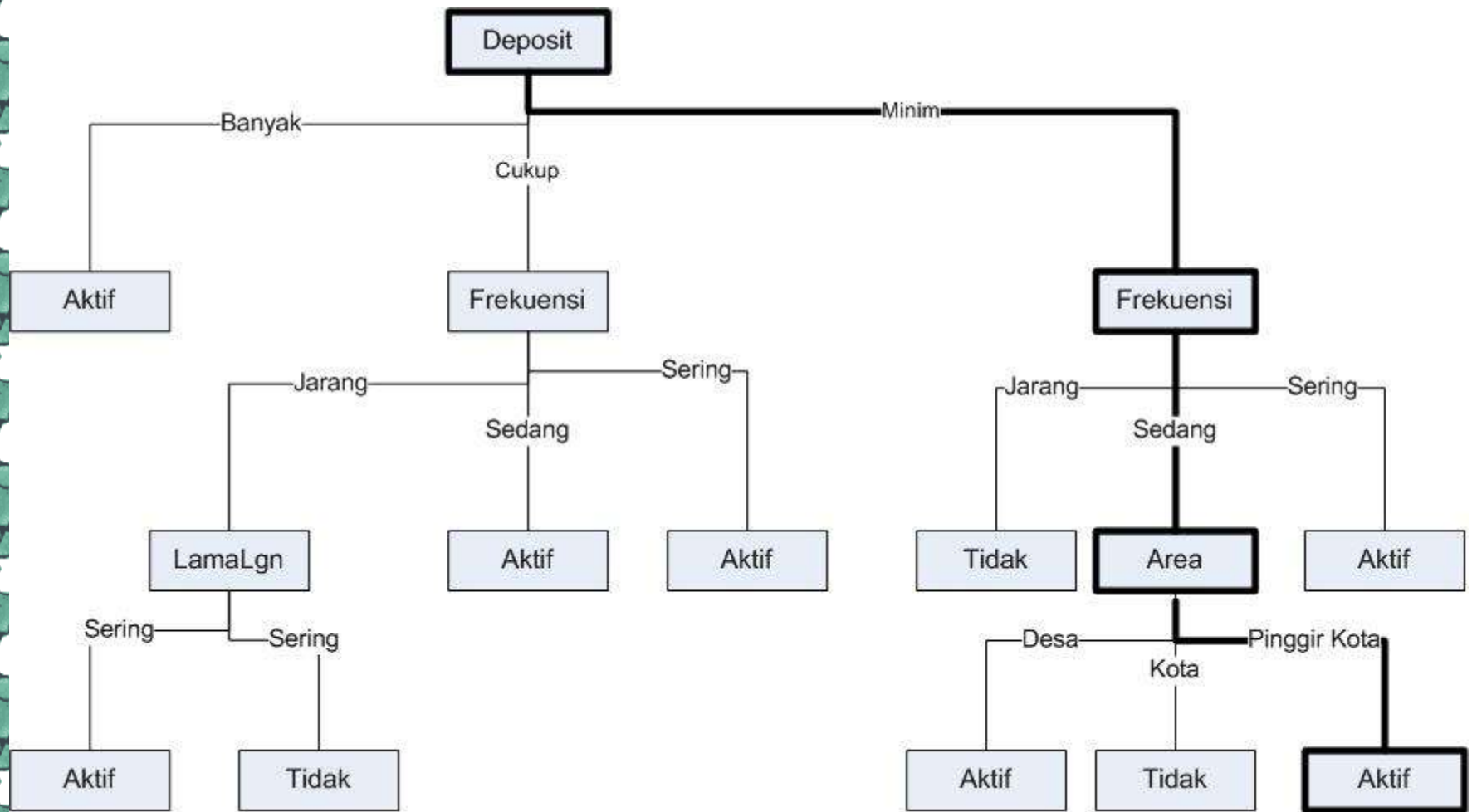
Model Pohon Keputusan hipotesa 6



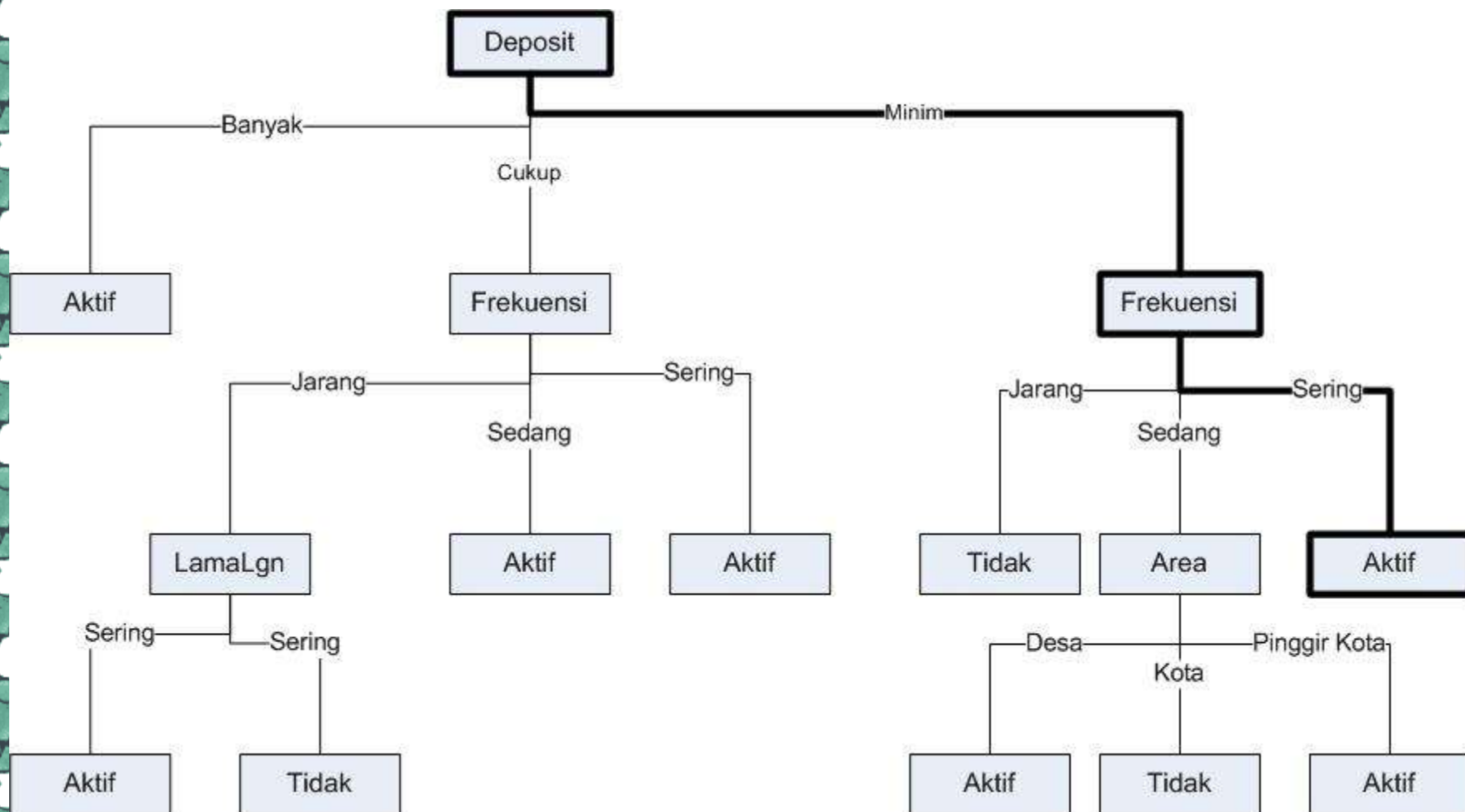
Model Pohon Keputusan hipotesa 7



Model Pohon Keputusan hipotesa 8



Model Pohon Keputusan hipotesa 9



Model Pohon Keputusan hipotesa 10

Bobot Nilai Iterasi 1

Hipotesa	Error (e_m)	α_m	W_m	hasil hipotesa
1	0	0	1	0
2	0.25	2.1972246	0.111111	0.244136064
3	0.25	2.1972246	0.111111	0.244136064
4	0.16	3.3164562	0.03628	0.12032494
5	0	0	1	0
6	0.25	2.1972246	0.111111	0.244136064
7	0.5	0	1	0
8	0	0	1	0
9	0.33	1.4163701	0.24259	0.343601483
10	0	0	1	0

Bobot Nilai Iterasi 2

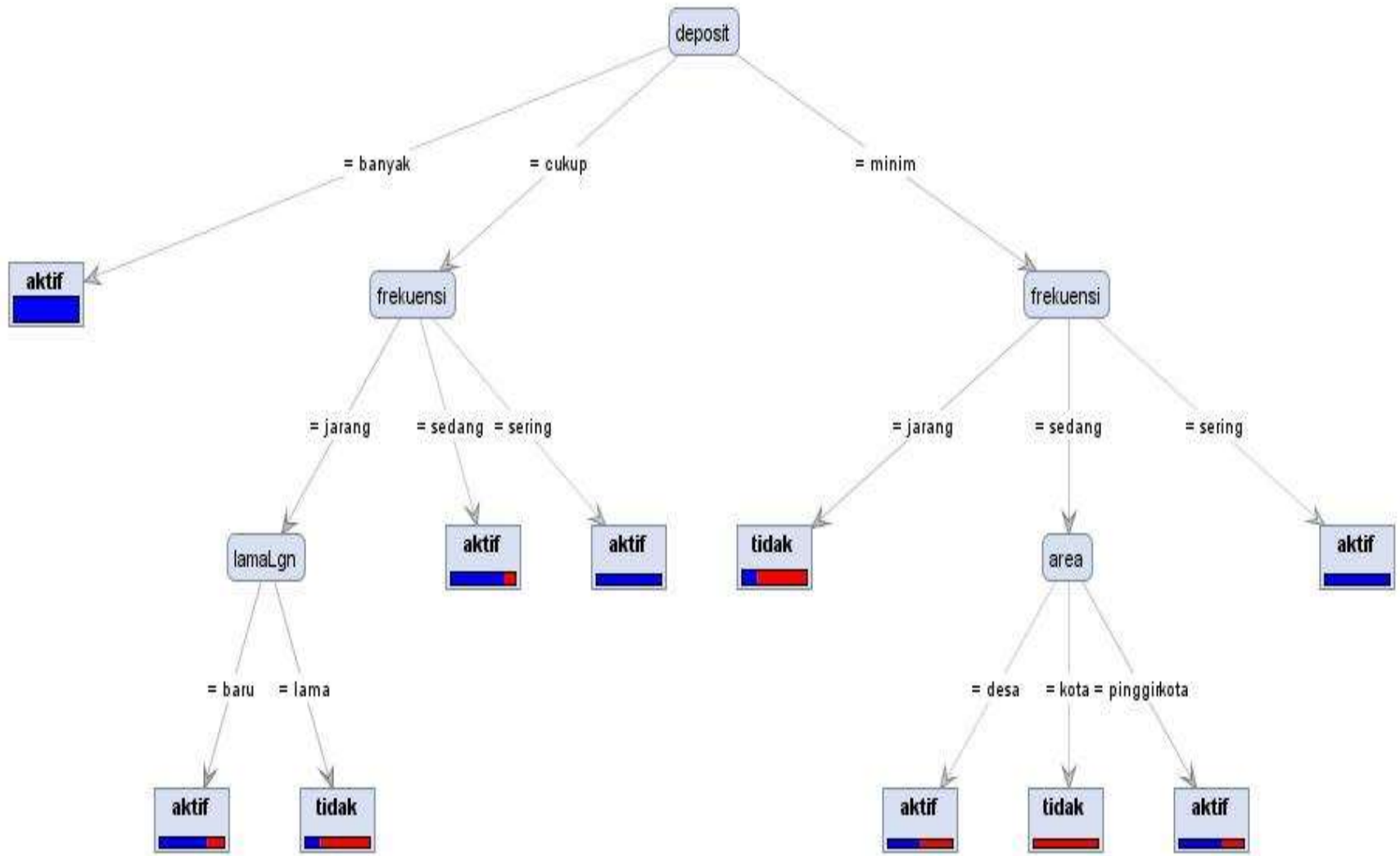
Hipotesa	Error (e_m)	α_m	W_m	hasil hipotesa
1	0	0	1	0
2	0.2441361	2.260271	0.104322	0.23579642
3	0.2441361	2.260271	0.104322	0.23579642
4	0.1203249	3.978713	0.01871	0.07444052
5	0	0	1	0
6	0.2441361	2.260271	0.104322	0.23579642
7	0	0	1	0
8	0	0	1	0
9	0.3436015	1.294571	0.274015	0.354732364
10	0	0	1	0

3. Menentukan nilai akhir class

Hasil α_m dari masing-masing hipotesa dari iterasi ke-1 dan iterasi ke-2 digabungkan didapatkan nilai akhir +1.

Hipotesa	α_m (iterasi 1)	α_m (iterasi 2)	sign ($\alpha_{m1} + \alpha_{m2}$)
2	2.197224577	2.260271342	1
3	2.197224577	2.260271342	1
4	3.316456153	3.978713349	1
6	2.197224577	2.260271342	1
7	0	2.260271342	1
9	1.416370116	1.294571119	1

Model pohon keputusan C4.5 berbasis AdaBoost





3. Evaluasi dan Validasi

1. Confusion matrix
2. Kurva ROC

Confusion Matrix

Akurasi Algoritma C4.5

accuracy: 71.33% +/- 18.15% (mikro: 71.19%)			
	true aktif	true tidak	class precision
pred. aktif	35	10	77.78%
pred. tidak	7	7	50.00%
class recall	83.33%	41.18%	

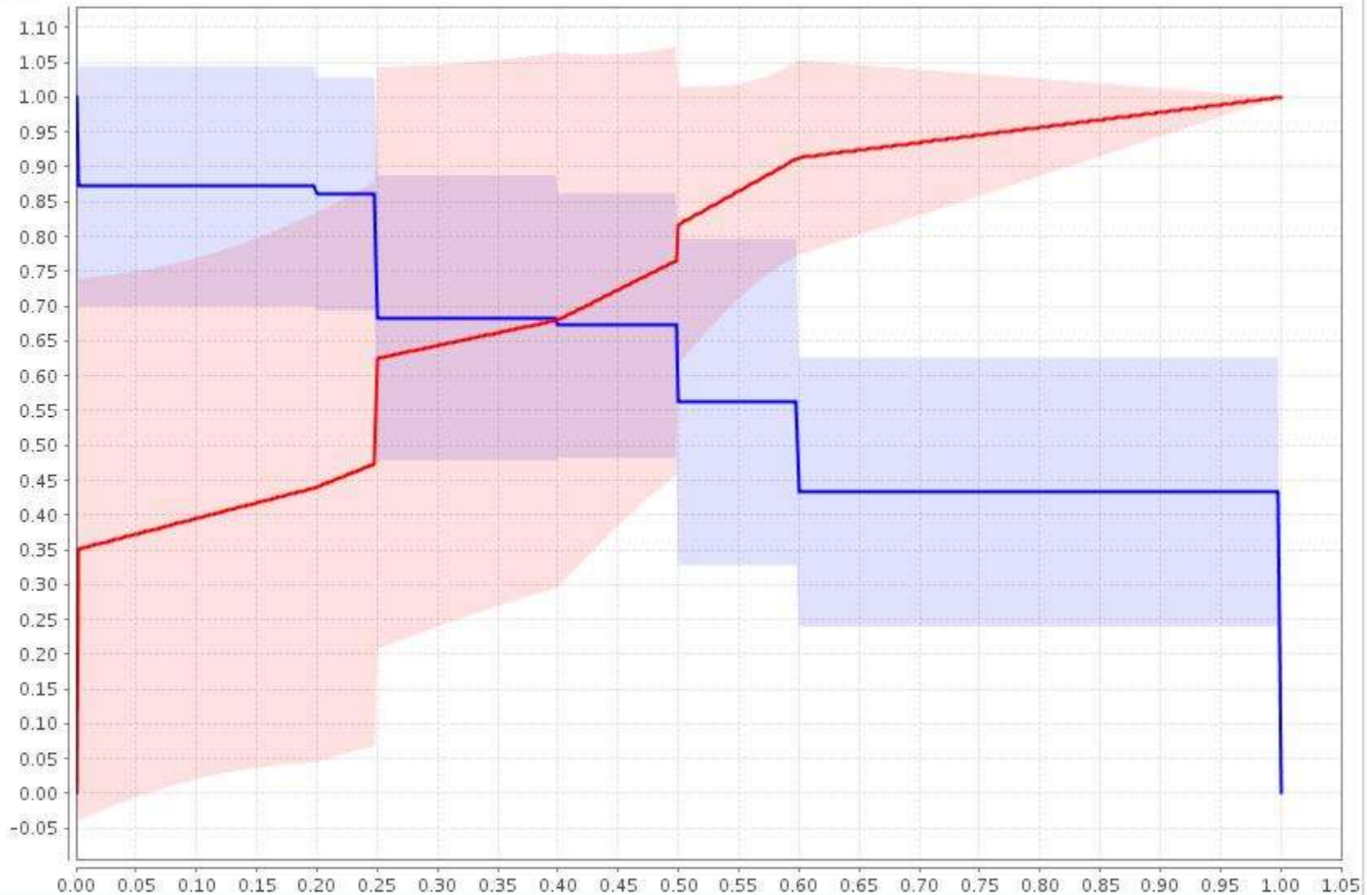
Akurasi C4.5 berbasis AdaBoost

accuracy: 88.00% +/- 7.92% (mikro: 88.14%)			
	true aktif	true tidak	class precision
pred. aktif	39	4	90.70%
pred. tidak	3	13	81.25%
class recall	92.86%	76.47%	

Kurva ROC – Algoritma C4.5

AUC: 0.741 +/- 0.196 (mikro: 0.741) (positive class: tidak)

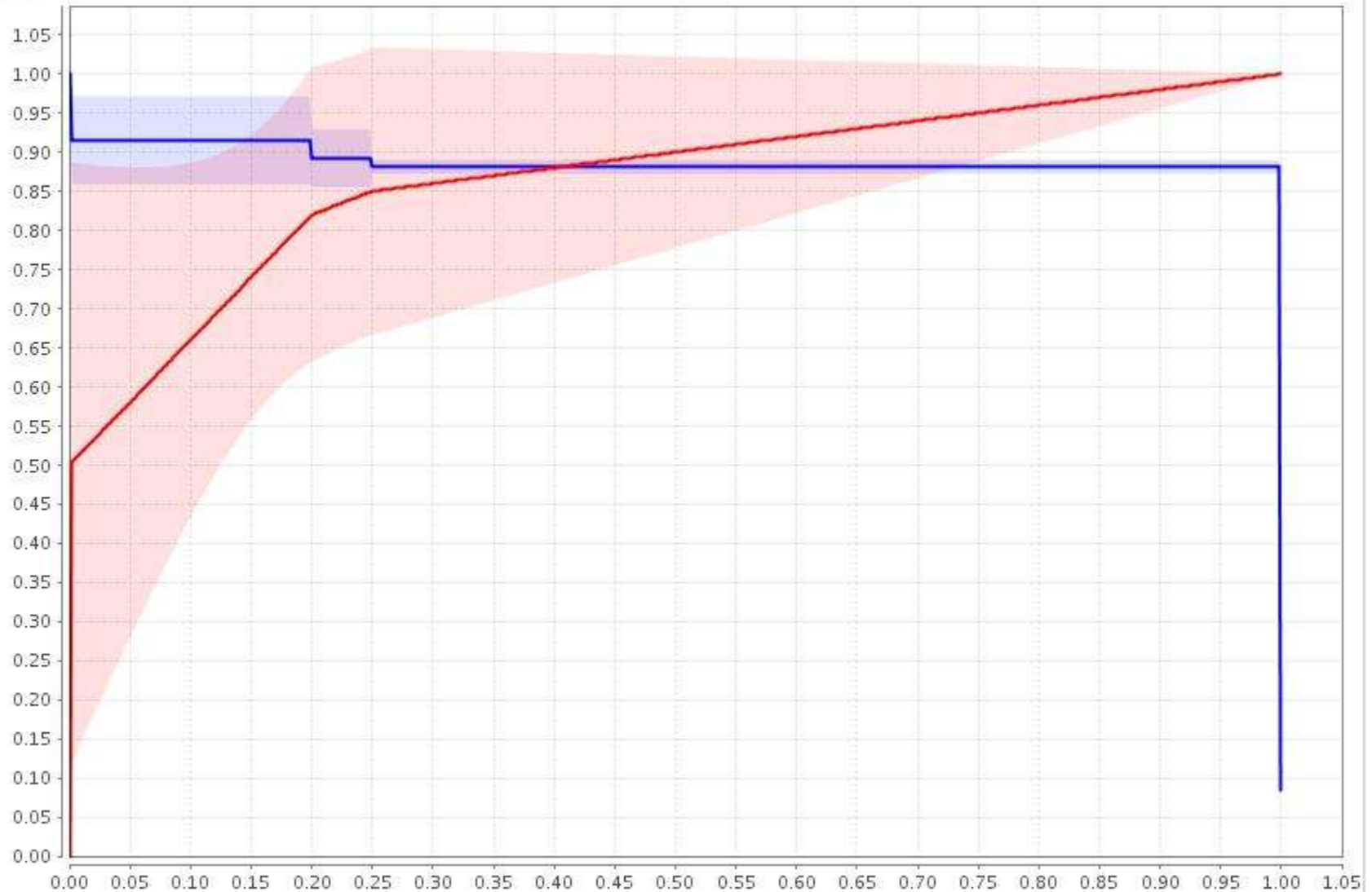
— ROC — ROC (Thresholds)



Kurva ROC – C4.5 berbasis AdaBoost

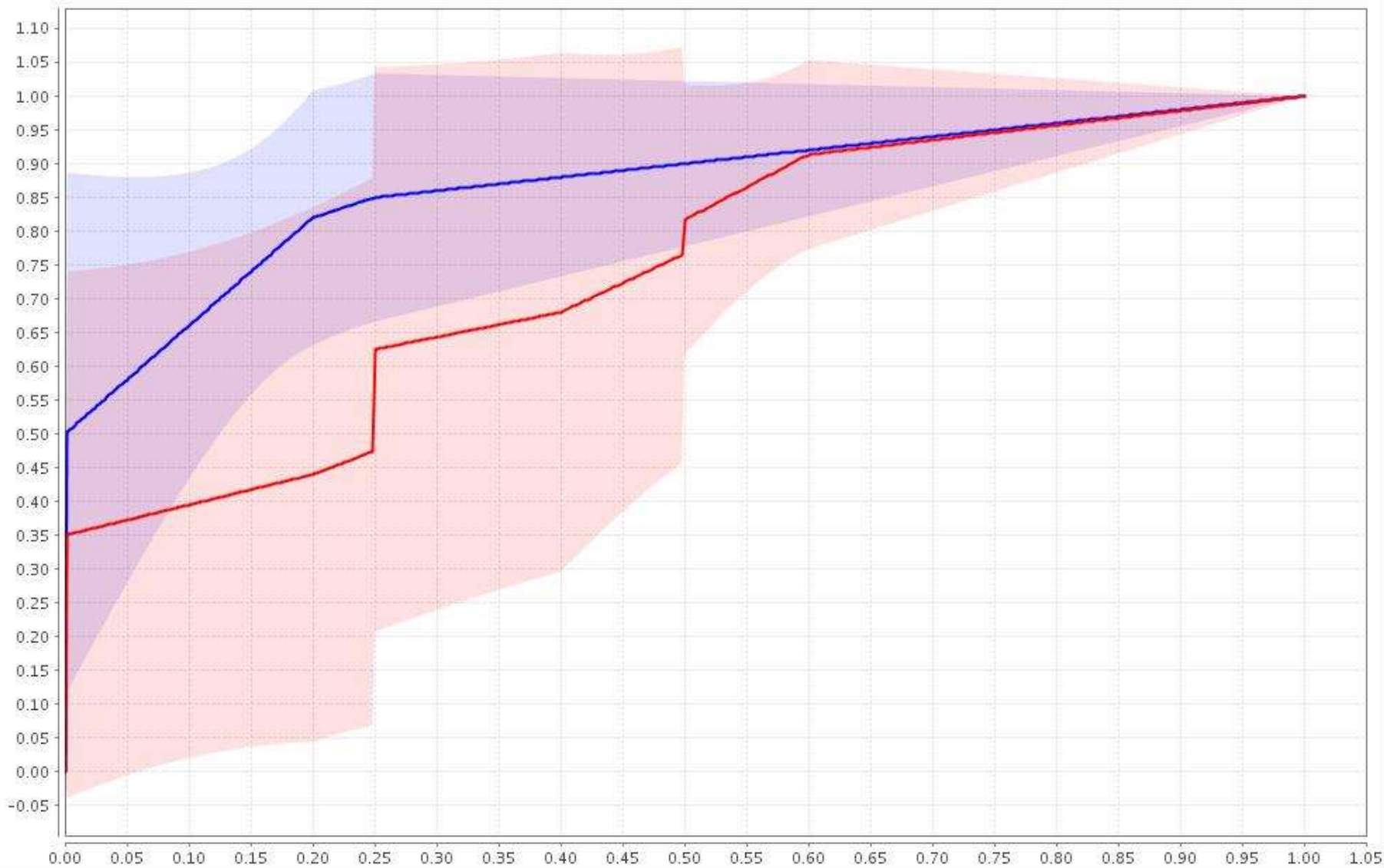
AUC: 0.868 +/- 0.105 (mikro: 0.868) (positive class: tidak)

— ROC — ROC (Thresholds)



Komparasi Kurva ROC

Decision Tree (2) AdaBoost



Analisa Hasil Komparasi

	Algoritma C4.5	C4.5 berbasis AdaBoost
Accuracy	71.33%	88.00%
AUC	0.741	0.868

- Penggunaan model C4.5 berbasis AdaBoost lebih akurat dalam memprediksi loyalitas pelanggan
- Hasil dari grafik AUC menunjukkan bahwa model C4.5 berbasis AdaBoost menunjukkan model tersebut mempunyai *good classification*

Implikasi Penelitian

- **Implikasi kebijakan** yang dirujuk kepada manajerial perusahaan:
 - Memberikan bonus atau potongan harga bagi pelanggan yang mempunyai **deposit** dalam jumlah banyak dan **frekuensi** aktivitas pengisian pulsa elektronik tinggi
 - Memberikan harga kompetitif bagi pelanggan yang tinggal di per**kota**an dimana tingkat persaingannya lebih tinggi
 - Memberikan pelayanan yang lebih cepat dan baik terutama pada pelanggan **lama**
- **Implikasi teoritis** yang di rujuk kepada pengembangan teoritis:
 - Untuk menghasilkan pola prediksi yang bervariasi, sebaiknya menggunakan metode Bagging

Kesimpulan

- Algoritma C4.5 berbasis AdaBoost meningkatkan akurasi dalam memprediksi pola dan perilaku loyalitas pelanggan
- Berdasarkan penelitian tingkat akurasi prediksi dengan model C4.5 berbasis AdaBoost sebesar 88% sedangkan algoritma C4.5 sebesar 71.33%.

Saran

- Dengan meningkatnya akurasi dalam memprediksi pola dan perilaku loyalitas pelanggan, perusahaan dapat memberikan pelayanan yang lebih baik kepada pelanggan
- Pada penelitian selanjutnya akan dikembangkan dalam menguji ulang model penelitian dengan menambah atribut yang lain seperti data keluhan pelanggan
- Untuk menambah ketelitian dan akurasi algoritma, akan lebih baik apabila metode AdaBoost dibandingkan dengan metode lain seperti Bayesian Boosting



Sekian dan Terimakasih