ALGORITMA C4.5 UNTUK PREDIKSI HASIL PEMILIHAN LEGISLATIF DPRD DKI JAKARTA

[1]Evicienna, [2]Hilda Amalia
[1,2]Jurusan Komputerisasi Akuntansi AMIK Bina Sarana Informatika Jakarta
Jl. Ciledug Raya No. 168 Ulujami, Jakrta Selatan
email: [1]evicienna.eca@bsi.ac.id, [2]hilda.ham@bsi.ac.id

ABSTRACT

Elections are a means of implementation of the sovereignty of the people in the Unitary State of Indonesia based on Pancasila and 1945 Constitution. Elections held in Indonesia is to choose the leadership of both the president and vice president, member of parliament, parliament, and the DPD. The election results should be predicted accurately, because an impact on various aspects of social, economic, security, and others. Based on these problems we need a model that can accurately predict the outcome of the election. C4.5 algorithm model is a model that is easy to understand and has a good degree of accuracy. By using the C4.5 algorithm models the obtained results are of high accuracy values for the election results in the amount of 97.84% and the AUC value obtained is 0.970 with a diagnosis rate Excellent Classification

Keyword: *C4.5 Algorithm, Prediciton, Election*

•

I. Pendahuluan

Pemilu adalah sarana pelaksanaan kedaulatan rakyat dalam Negara Kesatuan RI yang berdasarkan pancasila dan UUD 1945(UU RI nomor 3 tahun 1999). Pemilihan Umum (pemilu) merupakan salah satu pilar utama untuk memilih pimpinan(Sardini, 2011). Untuk penetapan caleg DPRD terpilih dilaksanakan dengan sistem suara terbanyak pada pemilu tahun 2009. Dengan ketentuan suara terbanyak, penetapan caleg terpilih ditetapkan peringkat suara sah terbanyak pertama, kedua, ketiga, dan seterusnya. Ketentuan ini tertuang dalam peraturan KPU nomor 15 tahun 2009. Ketentuan ini membuat sistem penetapan calon terpilih menjadi berbeda dengan pemilu tahun 2004. Setiap lima tahun tatacara perhitungan suara selalu berubah sesuai dengan peraturan perundang-undangan yang berlaku. Prediksi hasil pemilihan umum perlu diprediksi dengan akurat, karena hasil prediksi yang akurat sangat penting dan mempunyai dampak diberbagai aspek sosial, ekonomi, keamanan, dan lain-lain(Borisyuk, Borisyuk, Rallings, & Thrasher, 2005). Bagi para pelaku ekonomi, peristiwa politik seperti pemilu tidak dapat dipandang sebelah mata, mengingat hal tersebut dapat mengakibatkan risiko positif maupun negatif terhadap kelangsungan usaha yang dijalankan.

Algoritma C4.5 atau disebut dengan pohon keputusan adalah sebuah pohon dimana terdapat node internal yang mendeskripsikan atribut-atribut, setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas. Pohon keputusan dengan mudah dapat dikonversi ke aturan klasifikasi. Secara umum pohon keputusan memiliki akurasi yang baik, namun keberhasilan penggunaan tergantung pada data yang diolah.

Penelitian sebelumnya pernah membahas mengenai prediksi pemilu dengan motode *decision tree* maupun *classification tree* dengan nilai akurasi baik. Dari beberapa penelitian tersebut dilakukan dengan model *clasification tree* dan *neural network* hasil akurasi yang didapat untuk *classification tree* sebesar 92,82% dana *neural network* sebesar 96,82%(Vishnuprasad, 2005), model *k-Feature Set* hasil akurasinya 80%(Moscato, Mathieson, & Berretta, 2004), dan model CHAID (*Chi-squared Automatic Interaction Detection*) total estimasi yang didapat sebesar 97,9%(Choi & Han, 1999).

Pada penelitian ini akan memprediksi hasil pemilu dengan menggunakan algoritma dengan menganalisis sejumlah atribut yang menjadi parameter untuk prediksi hasil pemilu DPRD DKI Jakarta, diantaranya: nama partai, no urut partai, suara sah partai, nama caleg, kota administrasi, jenis kelamin, suara sah caleg, no urut caleg, jumlah perolehan kursi, dan daerah pemilihan.

II. Kajian Literatur

a. Data Mining

Data mining adalah salah satu cabang ilmu komputer yang banyak menarik perhatian masyarakat. Data mining adalah tumpukan data yang sudah tersimpan selama bertahun-tahun yang dimasukkan kedalam database tetapi tidak digunakan kembali atau disebut "data sampah". Data mining digunakan untuk menggali dan mendapatkan informasi dari data dengan jumlah besar(Gorunescu, 2011). Salah satu metode data mining adalah pengklasifikasian data, yaitu kegiatan mengekstrak dan memprediksi label kategori untuk masing-masing data. Adapun salah satu model pengklasifikasin dari data mining tersebut adalah algoritma C4.5 yang akan dijadikan sebagai model pada penelitian ini.

b. Algoritma C4.5

Algoritma C4.5 adalah hasil dari pengembangan algoritma ID3 (*Iterative Dichotomiser*) yang dikembangkan oleh Quinlan(Han & Kamber, 2006). Algoritma C4.5 atau pohon keputusan mirip sebuah pohon dimana terdapat node internal (bukan daun) yang mendeskripsikan atribut-atribut, setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas. Pohon keputusan dengan mudah dapat dikonversi ke aturan klasifikasi. Secara umum keputusan pengklasifikasi pohon memiliki akurasi yang baik, namun keberhasilan penggunaan tergantung pada data yang akan diolah.

Adapun tahapan yang digunakan dalam membuat sebuah pohon keputusan menggunakan algoritma C4.5 yang ada di penelitian ini(Gorunescu, 2011) yaitu:

1. Mempersiapkan data *training*, dapat diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokan dalam kelas-kelas tertentu.

2. Menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari masing-masing atribut atau berdasarkan nilai *index entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai *index entropy*, dengan rumus:

$$Entropy\,(i) = -\sum_{j=1}^{m} f(i,j).\log 2\, f[(i,j)]$$

Keterangan:
i = himpunan kasus
m = jumlah partisi i
f(i,j) = proposi j terhadap i

3. Hitung nilai *gain* dengan rumus:

$$Entropy\ split = -\sum_{i=1}^{p} \frac{n1}{n}\ .\ IE\,(i)$$

Keterangan:
p = jumlah partisi atribut
ni = proporsi ni terhadap i
n = jumlah kasus dalam n

4. Ulangi langkah ke-2 hingga semua *record* terpartisi

Proses partisi pohon keputusan akan berhenti disaat:

a. Semua tupel dalam *record* dalam simpul m mendapat kelas yang sama
b. Tidak ada atribut dalam *record* yang dipartisi lagi
c. Tidak ada *record* didalam cabang yang kosong.

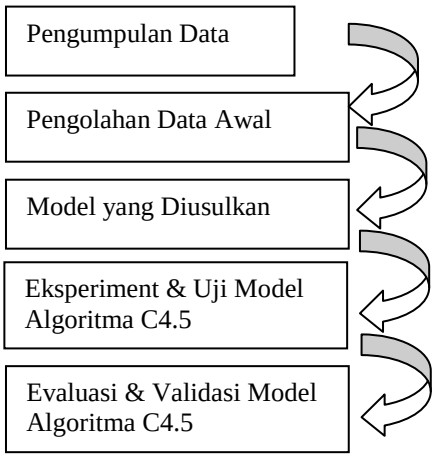
III. Metode Penelitian

Terdapat empat metode penelitian yang umum digunakan, yaitu A*ction Research*, E*xperiment*, C*ase Study*, dan S*urvey*(Dawson, 2009). Pada penelitian kali ini yang digunakan adalah penelitian *Experiment*, yaitu penelitian yang melibatkan penyelidikan perlakuan pada parameter/variabel tergantung dari penelitinya dan menggunakan tes yang dikendalikan oleh si peneliti itu sendiri. Dalam penelitian eksperimen digunakan spesifikasi *software* dan *hardware* sebagai alat bantu dalam penelitian pada Tabel 1.

Tabel 1 Spesifikasi *Hardware* dan *Software*

Software	Hardware
Sistem Operasi: Win 7	CPU: Dual Core
Data Mining: Rapid Miner	Memory: 1 GB
	Hardisk; 250 GB

Pada penelitian ini, data yang digunakan adalah data pemilu tahun 2009. Data pemilu tersebut akan diolah menggunakan model algoritma C4.5 untuk mendapatkan nilai akurasi yang baik dan dapat digunakan sebagai *rules* dalam memprediksi hasil pemilu. Dalam penelitian ini akan dilakukan beberapa langkah-langkah atau tahapan penelitian seperti yang terdapat pada gambar 1.



Gambar 1 Tahapan Penelitian

1. Pengumpulan Data

Pada penelitian ini digunakan pengumpulan data sekunder, yaitu mengambil data pemilu ditahun 2009, menggunakan buku, jurnal, publikasi, dan lain-lain. Data yang didapat dari KPUD Jakarta adalah data pemilu tahun 2009 dengan jumlah data sebanyak 2268 *record*, terdiri dari 11 variabel atau atribut. Adapaun variabel prediktor yaitu no urut partai, nama partai, suara sah partai, no urut caleg, nama caleg, jenis kelamin, kota adminstrasi, daerah pemilihan, suarah sah caleg, jumlah perolehan kursi. Sendangkan variabel tujuannya yaitu hasil pemilu. Berikut contoh data pemilu tahun 2009 pada Tabel 2:

Tabel 2. Data Pemilu Tahun 2009

Nama Partai	Nama Caleg Lengkap	JK	Kota Administrasi	No Urut Parpol	Suara Sah Partai	Jumlah Perolehan Kursi	Daerah Pemilihan	No Urut Caleg	Suara Sah Caleg	Hasil Pemilu
Partai Hati Nurani Rakyat	H. Jamaluddin Lamanda, SH	L	Kota Administrasi Jakarta Utara	1	20917	6	DP-1	1	2425	TIDAK
Partai Hati Nurani Rakyat	Suprawito	L	Kota Administrasi Jakarta Utara	1	20917	6	DP-1	2	3348	YA
Partai Karya Peduli Bangsa	Drs. Haris	L	Kota Administrasi Jakarta Utara	2	3255	0	DP-1	1	385	TIDAK
Partai Karya Peduli Bangsa	Ir. Thomas Julius Soputan	L	Kota Administrasi Jakarta Utara	2	3255	0	DP-1	2	302	TIDAK
Partai Pengusaha Dan Pekerja Indonesia	Eddy Pardede, SE	L	Kota Administrasi Jakarta Timur	3	3801	0	DP-3	1	1259	TIDAK
Partai Pengusaha Dan Pekerja Indonesia	Pintor Pooma Gurning	L	Kota Administrasi Jakarta Timur	3	3801	0	DP-3	2	241	TIDAK
Partai Peduli Rakyat Nasional	Achmad Bachtiar, SH	L	Kota Administrasi Jakarta Selatan	4	2888	0	DP-4	2	310	TIDAK
Partai Gerakan Indonesia Raya	Muhammad Syaiful Jihad	L	Kota Administrasi Jakarta Selatan	5	35464	6	DP-1	4	1909	TIDAK
Partai Gerakan Indonesia Raya	Ir. S. Andyka	L	Kota Administrasi Jakarta Utara	5	35464	6	DP-1	5	4014	YA
Dan Seterusnya										

2. Pengolahan Data Awal

Untuk mendapatkan data yang berkualitas terdapat teknik *preprocessing* yang digunakan pada penelitian ini, yaitu:

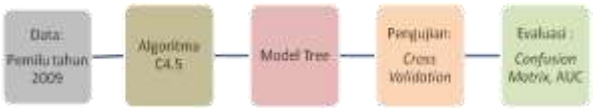
1. *Data integration and transformation*, untuk meningkatkan akurasi dan efisiensi algoritma(Vercellis, 2009). Data yang digunakan dalam penulisan ini bernilai kategorikal. Data ditransformasikan kedalam angka menggunakan *software* RapidMiner, terlihat pada Tabel 3

Tabel 3 Candidate Split Algoritma C4.5

Candidate Split	Child Nodes	
1	Suara sah caleg ≤ 6989.500	Suara sah caleg > 6989.500
	Suara sah caleg ≤ 4295.500	Suara sah caleg > 4295.500
	Suara sah caleg ≤ 8529.500	Suara sah caleg > 8529.500
	Suara sah caleg ≤ 4488	Suara sah caleg > 4488
	Suara sah caleg ≤ 2589.500	Suara sah caleg > 2589.500
	Suara sah caleg ≤ 2919.500	Suara sah caleg > 2919.500
	Suara sah caleg ≤ 5600.500	Suara sah caleg > 5600.500
	Suara sah caleg ≤ 5898	Suara sah caleg > 5898
	Suara sah caleg ≤ 6319	Suara sah caleg > 6319
	Suara sah caleg ≤ 3643.500	Suara sah caleg > 3643.500
2	Suara sah partai ≤ 49839	Suara sah partai > 49839
	Suara sah partai ≤ 16982.500	Suara sah partai > 16982.500
	Suara sah partai ≤ 43039.500	Suara sah partai > 43039.500
	Suara sah partai ≤ 46526	Suara sah partai > 46526
	Suara sah partai ≤ 21615	Suara sah partai > 21615
	Suara sah partai ≤ 28368.500	Suara sah partai > 28368.500
	Suara sah partai ≤ 26112	Suara sah partai > 26112
3	Jumlah perolehan kursi ≤ 14.500	Jumlah perolehan kursi > 14..500
	Jumlah perolehan kursi ≤ 5	Jumlah perolehan kursi > 5
4	Nama partai = Partai Amanat Nasional	
	Nama partai = Partai Damai Sejahtera	
	Nama partai = Partai Demokrasi Indonesia Perjuangan	
	Nama partai = Partai Demokrat	
	Nama partai = Partai Gerakan Indonesia Raya	
	Nama partai = Partai Golkar	
	Nama partai = Partai Keadilan Sejahtera	
	Nama partai = Partai Peduli Rakyat Nasional	
	Nama partai = Partai Persatuan Pembangunan	

3. Model yang Diusulkan

Model yang diusulkan pada penelitian ini adalah menggunakan algoritma C4.5, yang terlihat pada Gambar 2



Gambar 2 Model yang Diusulkan

Algoritma C4.5 yaitu model untuk mengubah data menjadi pohon keputusan dengan aturan-aturannya (*rules*).

IV. PEMBAHASAN

a. Eksperimen

Adapun eksperimen yang dilakukan pada penelitian ini adalah:

1. Menghitung jumlah kasus class YA dan class TIDAK serta nilai *Entropy* dari semua kasus. Kasus dibagi berdasarkan atribut pada Tabel 2.3 dengan jumlah kasus 2268 *record*, kelas YA ada 94 *record* dan kelas TIDAK sebanyak 2174 *record* sehingga didapat *entropy*:

$$Entropy\,(i) = -\sum_{j=1}^{m} f(i,j).\log 2\, f[(i,j)]$$

= (-94/2268 . $\log_2$ (94/2268)) + (-2174/2268 . $\log_2$ (2174/2268))
= 0.2488

2. Hitung nilai Gain dari masing-masing atribut pada Tabel 3, sebagai contoh untuk suara sah caleg:

≤ 6989.500 = 2203/2268
> 6989.500 = 65/2268
Atribut suara sah caleg ≤ 6989.500 terdiri dari 32 class YA dan 2171 class TIDAK, dan untuk atribut suara sah caleg > 6989.500 terdiri dari 62 class YA dan 3 class TIDAK. Nilai Entropynya dapat dihitung sebagai berikut:

$$Entopy\ split = \sum_{i=1}^{p} \frac{ni}{n}\ IE\,(i)$$

Suara sah caleg ≤ 6989.500 = ((-32/2203 . log2 (32/2203) + (-2171/2203. log2 (2171/2203))
= 0.10949

Suara sah caleg > 6989.500 = ((-62/65 . log2 (62/65) + (-3/65 . log2 (3/65))
= 0.26983

E split suara sah caleg = ((2203/2268 (0.10949) + (65/2268 (0.26983))
= 0.11408

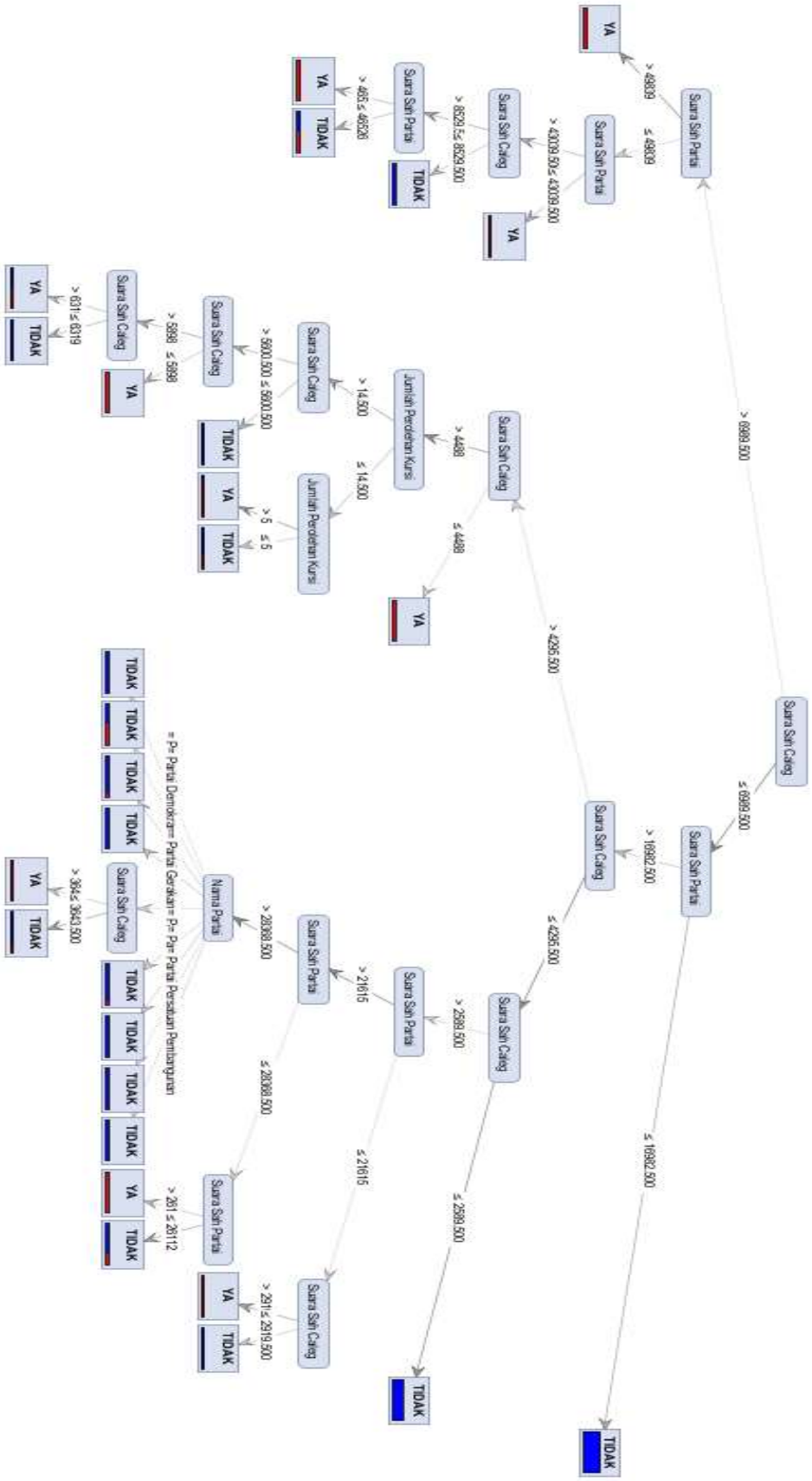
Gain suara sah caleg = 0.2488 – 0.11408
= 0.1347

Perhitungan *entropy* dan *gain* untuk semua atribut dilakukan, untuk mendapatkan nilai gain tertinggi. Hasil perhitungan seluruh atribut terlihat pada tabel 4

Tabel 4 Informasi Gain Algoritma C4.5

Candidate Split		Kasus	YA	TIDAK	Entropy	Gain
suara sah caleg						
<=	6989.500	2203	32	2171	0.114081	0.134719
>	6989.500	65	62	3		
<=	4295.500	2162	14	2148	0.091322	0.157478
>	4295.500	106	80	26		
<=	8529.500	2220	47	2173	0.147927	0.100873
>	8529.500	48	47	1		
<=	4488	2166	18	2148	0.103082	0.145718
>	4488	102	76	26		
<=	2589.500	2055	0	2055	0.09298	0.15582
>	2589.500	213	94	119		
<=	2919.500	2078	2	2076	0.093822	0.154978
>	2919.500	190	92	98		
<=	5600.500	2185	23	2162	0.102994	0.145806
>	5600.500	83	71	12		
<=	5898	2190	27	2163	0.112752	0.136048
>	5898	78	67	11		
<=	6319	2198	29	2169	0.109622	0.139178
>	6319	70	65	5		
<=	3643.500	2132	11	2121	0.101675	0.147125
>	3643.500	136	83	53		
suara sah partai						
<=	49839	1802	23	1779	0.204852	0.043948
>	49839	466	71	395		
<=	16982.500	1338	0	1338	0.193707	0.055093
>	16982.500	930	94	836		
<=	43039.500	1758	20	1738	0.203941	0.044589
>	43039.500	510	74	436		
<=	46526	1778	22	1756	0.205463	0.043337
>	46526	490	72	418		

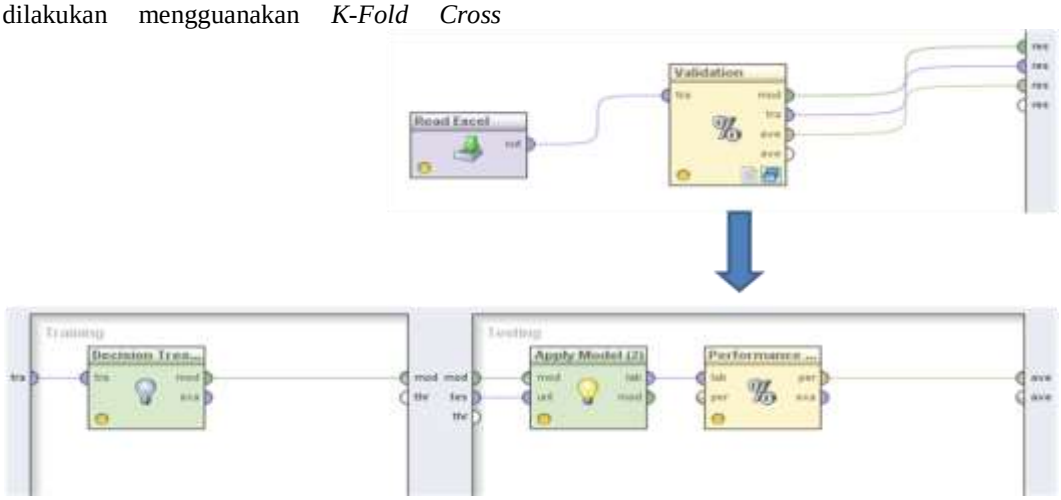
<=	21615	1452	5	1447	0.182026	0.066774
>	21615	816	89	727		
<=	28368.500	1571	12	1559	0.2055406	0.043394
>	28368.500	697	82	615		
<=	26112	1530	10	1520	0.204726	0.044074
>	26112	738	84	654		
jumlah perolehan kursi						
<=	14.500	2053	44	2009	0.209417	0.039383
>	14.500	215	50	165		
<=	5	1535	9	1526	0.202427	0.046373
>	5	733	85	648		
nama partai						0.085542
PAN		109	4	105	0.010907	
PDS		98	4	94	0.010631	
PDIP		111	11	100	0.022813	
Partai Demokrat		112	32	80	0.042623	
Gerindra		74	6	68	0.013246	
Partai Golkar		112	7	105	0.016656	
PKS		101	18	83	0.030111	
PPRN		85	0	85	0	
PPP		103	7	96	0.016271	



Gambar 3 Model Pohon Keputusan Algoritma C4.5

b. Pengujian

Sementara untuk pengujian yang dilakukan mengguanakan *K-Fold Cross Validation* yang terlihat pada gambar dibawah ini:



Gambar 4 Pengujian K-Fold Cross Validation Algoritma C4.5

Model klasifikasi bisa dievaluasi berdasarkan kriteria seperti tingkat akurasi, kecepatan, kehandalan, skabilitas, dan interpretabilitas(Vercellis, 2009). Mencari nilai akurasi dan AUC menggunakan model *confusion matrix* dan ROC (*Receiver Operating Characteristic*).

a. *Confusion Matrix*

Berikut tabel *Confusion Matrix* algoritma C4.5, dari tabel ini terlihat nilai akurasi yang didapat adalah 97,84%.

Tabel 5 Model *counfusion matrix* untuk Algoritma C4.5

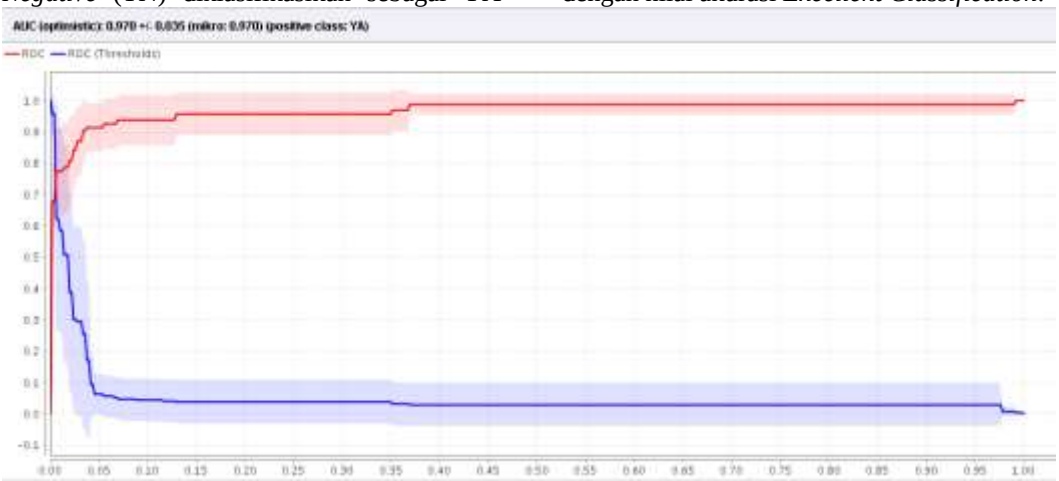
accuracy: 97.84% +/- 0.96% (mikro: 97.84%)

	true TIDAK	true YA	class precision
pred. TIDAK	2153	28	98.72%
pred. YA	21	66	75.86%
class recall	99.03%	70.21%	

Dari tabel diatas terlihat bahwa jumlah *True Positive* adalah 2153 *record* diklasifikasikan sebagai TIDAK terpilih dan *False Negative* sebanyak 28 *record* diklasifikasikan sebagai TIDAK terpilih tetapi YA terpilih. Berikutnya 66 *record* untuk *True Negative* (TN) diklasifikasikan sebagai YA terpilih, dan 21 *record False Positive* (FP) diklasifikasin sebagai YA terpilih ternyata TIDAK.

b. Evaluasi ROC

Pada grafik ROC terlihat nilai AUC(*Area Under Curve*) sebesar 0.970 dengan nilai akurasi *Excellent Classification*.



Gambar 5 Nilai AUC dalam grafik ROC algoritma C4.5

V. KESIMPULAN

Hasil eksperimen dan evaluasi penelitian prediksi hasil pemilihan legislatif DPRD DKI Jakarta menggunakan algoritma C4.5 terbukti akurat, terlihat dari hasil yang didapat yaitu sebesar nilai akurasi sebesar 97.84% dan nilai AUC sebesar 0.970 dengan tingkat diagnosa *Excellent Classification*.

DAFTAR PUSTAKA

Borisyuk, R., Borisyuk, G., Rallings, C., & thrasher, M. (2005). Forcesting the 2005 General Election: A Neural Network Approach. The British Journal of Politics & International Relations Volume 7, Issue 2, 145-299.

Choi, J. H., & Han, S. T. (1999). Predictionof Election Results Using Discrimination of Non-Respondents.

Dawson, C. W. (2009). Project in Computing and Information System A Student's Guide. England: Addison-Wesley.

Gorunescu, F. (2011). Data Mining Concepts, Model and Technique. Berlin: Springer.

Han, J., & Kamber, M. (2006). Data Mining Concepts and technique. San Francisco: Diane Cerra

Moscato, P. Mathieson, L. Mendes, A., & Berretta, R. (2004). The Electronic Primaries: Predicting the U.S. Presidency Using Feature Selection with Safe Data Reduction.

Sardini, N. H. (2011). Restorasi Penyelenggaraan Pemilu di Indonesia. Yogyakarta: Fajar Media Press

Undang-Undang Republik Indonesia Nomor 3 Tahun 1999. Tentang: Pemilihan Umum.

Vercellis, C. (2009). Business Inelligentce: Data Mining and Optimization for Decision Making. Southern Gate, Chichester, West Sussex: John Wiley & Sons, Ltd.

Vishnuprasad, N. (2005). Building Predictive Models for Election Results In Indiana-an Application of Classification