PREDIKSI HASIL PEMILU LEGISLATIF DENGAN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR

Mohammad Badrul
Program Studi Sistem Informasi
STMIK Nusa Mandiri Jakarta
Jl. Damai No. 8 Warung Jati Barat (Margasatwa) Jakarta Selatan.
Telp. (021) 78839513 Fax. (021) 78839421
mohammad.mbl@nusamandiri.ac.id

Abstract — *General Elections in Indonesia has undergone several changes of election period to another period Elections . During the New Order elections, we know the proportional electoral system with closed lists . The election of candidates is not determined voters , but the authority of elite political party in accordance with the order of the list of candidates along with the serial number . In such systems , political parties become very strong position against its cadres in parliament . But on the one hand , and the social basis of political relations representatives and constituents to be weak . This is what causes the position of elected candidates they become " distant " in conjunction with constituents . Excitement choose direct representatives of the people are just starting to be accommodated in the 2004 elections through Law no. 12 In 2003 , using a proportional system with open lists of candidates . Voters not only chose the symbol of a political party , but also given the opportunity to choose caleg.Penelitian relating to the election had been conducted by the researchers by using the decision tree method and classification tree and Bayesian estimators . In this study, researchers will use the K-Nearest Neighbor method . K-Nearest Neighbor have shown promising results in the prediction of time-series data as compared to the traditional approach so that the results of the legislative election predictions are more accurate Jakarta.*

Intisari — Pemilihan Umum di Indonesia telah mengalami beberapa perubahan dari periode Pemilu ke periode Pemilu yang lain. Selama pemilu Orde Baru, kita mengenal sistem pemilu proporsional dengan daftar tertutup. Keterpilihan calon legislatif bukan ditentukan pemilih, melainkan menjadi kewenangan elite partai politik sesuai dengan susunan daftar caleg beserta nomor urut. Dalam sistem demikian, kedudukan parpol menjadi sangat kuat terhadap kadernya di parlemen. Namun di satu sisi, basis sosial dan relasi politik para wakil rakyat dengan konstituen menjadi lemah. Inilah yang menyebabkan kedudukan caleg terpilih mereka menjadi ”jauh” dalam hubungannya dengan konstituen. Semangat memilih langsung wakil rakyat baru mulai diakomodasi pada Pemilu 2004 melalui UU No. 12 Tahun 2003, dengan menggunakan sistem proporsional dengan daftar calon terbuka. Pemilih tidak hanya memilih tanda gambar parpol, tetapi juga diberi kesempatan memilih caleg.Penelitian yang berhubungan dengan pemilu sudah pernah dilakukan oleh peneliti yaitu dengan menggunakan metode *decision tree* dan classification tree dan estimator bayesian. Pada penelitian ini peneliti akan menggunakan metode *K-Nearest Neighbor.* algoritma *K-Nearest Neighbor* telah menunjukkan hasil yang menjanjikan dalam prediksi untuk data time-series dibandingkan dengan pendekatan tradisional sehingga hasil prediksi pemilu legislatif DKI Jakarta lebih akurat.

Kata Kunci*:* akurasi, algoritma *K-Nearest Neighbor*, Pemilu.

PENDAHULUAN

Pemilihan Umum di Indonesia telah mengalami beberapa perubahan dari periode Pemilu ke periode Pemilu yang lain. Selama pemilu Orde Baru, kita mengenal sistem pemilu proporsional dengan daftar tertutup. Keterpilihan calon legislatif bukan ditentukan pemilih, melainkan menjadi kewenangan elite partai politik sesuai dengan susunan daftar caleg beserta nomor urut(Undang-Undang RI No.10 Tahun 2008). Dalam sistem demikian, kedudukan parpol menjadi sangat kuat terhadap kadernya di parlemen. Namun di satu sisi, basis sosial dan relasi politik para wakil rakyat dengan konstituen menjadi lemah. Inilah yang menyebabkan kedudukan caleg terpilih mereka menjadi ”jauh” dalam hubungannya dengan konstituen.

Sistem pemilu demikian juga dianggap membuat lembaga perwakilan rakyat menjadi elitis, ekslusif, tidak tersentuh oleh masyarakat, serta tidak sensitif terhadap problem rakyat. Seiring tuntutan reformasi tahun 1998, sistem pemilu tersebut mulai ditinggalkan. Pada Pemilu

1999, sistem yang digunakan pada dasarnya tidak mengalami perubahan dibandingkan pemilu Orde Baru dengan menggunakan sistem proporsional berdasarkan daftar tertutup. Pemilih masih terbatas mencoblos tanda gambar parpol.

Semangat memilih langsung wakil rakyat baru mulai diakomodasi pada Pemilu 2004 dengan menggunakan sistem proporsional dengan daftar calon terbuka. Pemilih tidak hanya memilih tanda gambar parpol, tetapi juga diberi kesempatan memilih caleg. Namun, penerapan ketentuan ini terkesan sengaja dilemahkan dengan pengaturan ketentuan suara sah dan penetapan calon terpilih. Suara sah parpol harus dicoblos bersamaan pada kolom tanda gambar parpol dan calegnya. Pemilih yang mencoblos caleg saja dianggap tidak sah. Sementara mencoblos tanda gambar parpol saja sah. Peraturan yang terkesan rumit dan tidak mempermudah pemilih untuk memilih caleg mereka secara langsung ini, dimanfaatkan oleh parpol dalam sosialisasi dan kampanye mereka untuk mencoblos tanda gambar parpol saja dengan dalih menghindari suara rusak atau tidak sah

Peraturan yang menggambarkan berlakunya sistem pemilu proporsional dengan daftar terbuka setengah hati ini masih dipersulit lagi dengan ketentuan penetapan caleg yang langsung terpilih, yang harus memenuhi ketentuan bilangan pembagi pemilihan (BPP). Jika tidak ada caleg yang memperoleh angka BPP, kursi yang didapat parpol di daerah pemilihan, menjadi hak caleg berdasarkan nomor urut terkecil. Sementara itu untuk mencapai angka BPP dengan membagi jumlah suara sah seluruh parpol peserta pemilu dengan jumlah kursi di daerah pemilihan (DPR, DPRD provinsi dan DPRD kabupaten/kota) sungguh sangat kecil kemungkinannya.

Pemilu bertujuan untuk memilih anggota DPR, DPRD provinsi, dan DPRD kabupaten/kota yang dilaksanakan dengan sistem proporsional terbuka[1]. Dengan sistem pemilu langsung dan jumlah partai yang besar maka pemilu legislatif memberikan peluang yang besar pula bagi rakyat Indonesia untuk berkompetisi menaikkan diri menjadi anggota legislatif. Pemilu legislatif tahun 2009 diikuti sebanyak 44 partai yang terdiri dari partai nasional dan partai lokal. Pemilu Legislatif DKI Jakarta Tahun 2009 terdapat 2.268 calon anggota DPRD dari 44 partai yang akan bersaing memperebutkan 94 kursi anggota Dewan Perwakilan Rakyat DKI Jakarta. Prediksi hasil pemilihan umum perlu diprediksi dengan akurat, karena hasil prediksi yang akurat sangat penting karena mempunyai dampak pada berbagai macam aspek sosial, ekonomi, keamanan, dan lain-lain. 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.

Metode prediksi hasil pemilihan umum sudah pernah dilakukan oleh peneliti melakukan prediksi hasil pemilihan umum dengan menggunakan metode *Estimator Bayesian*(Rigdon, Jacobson, Sewell, dan Rigdon, 2009), *dan* (Nagadevara dan Vishnuprasad, 2005) memprediksi hasil pemilihan umum dengan model *classification tree dan neural network*.

Pada penelitian ini peneliti akan menggunakan metode K-Nearest Neighbor untuk meprediksi hasil pemilu Legislatif DKI Jakarta sehingga hasil prediksi ini bisa digunakan oleh pihak yang terkait dengan Pemilihan tersebut.

BAHAN DAN METODE

A. Pemilihan Umum

Pemilihan umum adalah salah satu pilar utama dari sebuah demokrasi, kalau tidak dapat yang disebut yang terutama (Sardini, 2011). Pemilu di Indonesia terbagi dari dua bagian, yaitu Pemilu orde baru yaitu Sistem pemilihannya dilakukan secara proporsional tidak murni, yang artinya jumlah penentuan kursi tidak ditentukan oleh jumlah penduduk saja tetapi juga didasarkan pada wilayah administrasi dan pemilu era reformasi yaitu dikatakan sebagai pemilu reformasi karena dipercapatnya proses pemilu di tahun 1999 sebelum habis masa kepemimpinan di pemilu tahun 1997. Terjadinya pemilu era reformasi ini karena produk pemilu pada tahun 1997 dianggap pemerintah dan lembaga lainnya tidak dapat dipercaya.

Sistem pemilihan DPR/DPRD berdasarkan ketentuan dalam UU nomor 10 tahun 2008 pasal 5 ayat 1 sistem yang digunakan dalam pemilihan legislatif adalah sistem proporsional dengan daftar terbuka, sistem pemilihan DPD dilaksanakan dengan sistem distrik berwakil banyak UU nomor 10 tahun 2008 pasal 5 ayat 2. Menurut UU No. 10 tahun 2008, Peserta pemilihan anggota DPR/D adalah partai politik peserta Pemilu, sedangkan peserta pemilihan anggota DPD adalah perseorangan. Partai politik peserta Pemilu dapat mengajukan calon sebanyak- banyaknya 120 persen dari jumlah kursi yang diperebutkan pada setiap daerah pemilihan demokratis dan terbuka serta dapat mengajukan calon dengan memperhatikan keterwakilan perempuan sekurang-kurangnya 30 %. Partai Politik Peserta Pemilu diharuskan UU untuk mengajukan daftar calon dengan nomor urut (untuk mendapatkan Kursi). Karena itu dari

segi pencalonan UU No.10 Tahun 2008 mengadopsi sistem daftar calon tertutup.

Tanggal 31 Maret 2008, menjadi awal dari perubahan sistem Pemilihan Umum di Indonesia. Pemerintah mengesahkan Undang Undang Nomor 10 Tahun 2008 tentang Pemilihan Umum Anggota Dewan Perwakilan Rakyat, Dewan Perwakilan Daerah, Dan Dewan Perwakilan Rakyat Daerah. Secara umum, diberlakukannya Undang Undang Nomor 10 Tahun 2008 mengakibatkan berubahnya sistem pemilu di Indonesia, dari sistem proporsional terbuka "setengah hati", menjadi sistem proporsional yang memberi harapan semangat pilih langsung.

Pasal 5 Ayat (1) UU 10 Tahun 2008 tentang Pemilu Legislatif, tidak tampak berbeda dengan Pemilu 2004, tetap mendasarkan pada prinsip proporsional atau perwakilan berimbang. Artinya, suatu daerah pemilihan diwakili sejumlah wakil yang didapat dari perolehan suara partai-partai politik peserta pemilu. Yang membedakan, ketentuan penetapan caleg terpilih yang diatur dalam Pasal 214 yang didasarkan pada sistem nomor urut setelah tidak ada caleg yang memperoleh sekurang-kurangnya 30% BPP. Sementara caleg yang memenuhi ketentuan memperoleh sekurang-kurangnya 30% lebih banyak dari jumlah kursi yang diperoleh parpol peserta pemilu, kursi diberikan kepada caleg yang memiliki nomor urut lebih kecil di antara caleg yang memenuhi ketentuan sekurang-kurangnya 30% dari BPP. Dengan demikian sistem proporsional terbuka yang digunakan pada Pemilu 2009, masih tetap menerapkan pembatasan ketentuan perolehan suara sekurang-kurangnya 30% BPP bagi caleg untuk langsung ditetapkan sebagai caleg terpilih.

Bila kita menerapkan Undang Undang Nomor 10 Tahun 2008 pada hasil Pemilu 2004, dari 550 anggota DPR yang terpilih, hanya 116 orang (21,1%) yang memperoleh suara terbanyak dan sekurang-kurangnya mencapai 30% BPP. Sementara yang lain, sebagian besar anggota DPR 434 orang (78.9%) terpilih karena nomor urut dalam daftar calon. Artinya, posisi dalam nomor urut daftar calon tetap menjadi faktor yang lebih utama dalam menentukan seorang calon terpilih.

Keputusan Mahkamah Konstitusi menyatakan bahwa pasal 214 huruf a, b, c, d ,dan e UU No 10/2008 tentang Pemilu anggota DPR, DPD, dan DPRD, bertentangan dengan UUD RI 1945. Selanjutnya, menyatakan pasal 214 huruf a, b, c,d, dan e UU No 10/2008 tidak mempunyai kekuatan hukum mengikat. Pertimbangan dari putusan ini di antaranya, ketentuan pasal 214 huruf a,b,c,d, dan e UU No 10/2008 yang menyatakan bahwa calon anggota legislatif terpilih adalah calon yang mendapat suara di atas 30 persen dari bilangan pembagian pemilu (BPP) atau menempati nomor urut lebih kecil, dinilai bertentangan dengan makna substantif dengan prinsip keadilan sebagaimana diatur dalam pasal 28 d ayat 1 UUD 1945.

Dengan keluarnya Keputusan Mahkamah Konstitusi tersebut, maka penetapan calon anggota legislatif pada Pemilu 2009 tidak lagi memakai sistem nomor urut dan digantikan dengan sistem suara terbanyak. Hal ini memunculkan berbagai respon dari berbagai kalangan dan dari berbagai sisi.

UU No.10 Tahun 2008 mengadopsi sistem proporsional dengan daftar terbuka. sistem proporsional merujuk pada formula pembagian kursi dan/atau penentuan calon terpilih, yaitu setiap partai politik peserta pemilu mendapatkan kursi proporsional dengan jumlah suara sah yang diperolehnya. Penerapan formula proporsional dimulai dengan menghitung bilangan pembagi pemilih (BPP), yaitu jumlah keseluruhan suara sah yang diperoleh seluruh partai politik peserta pemilu pada suatu daerah pemilihan dibagi dengan jumlah kursi yang diperebutkan pada daerah pemilihan tersebut.

B. Data Mining

Data mining telah menarik banyak perhatian dalam dunia sistem informasi dan dalam masyarakat secara keseluruhan dalam beberapa tahun terakhir, karena ketersediaan luas dalam jumlah besar data dan kebutuhan segera untuk mengubah data tersebut menjadi informasi yang berguna dan pengetahuan. Informasi dan pengetahuan yang diperoleh dapat digunakan untuk aplikasi mulai dari pasar analisis, deteksi penipuan, dan retensi pelanggan, untuk pengendalian produksi dan ilmu pengetahuan eksplorasi (Han dan Kamber, 2007). Adanya ketersediaan data yang melimpah, kebutuhan akan informasi atau pengetahuan sebagai sarana pendukung dalam pengambilan keputusan baik bagi individu, organisasi, perusahaan dan pemerintahan.

Banyaknya data, ditambah dengan kebutuhan untuk alat analisis data yang kuat, telah digambarkan sebagai kaya data tapi miskin informasi. Jumlah data yang tumbuh secara cepat, dikumpulkan dan disimpan dalam repositori data yang besar dan banyak, telah jauh melampaui kemampuan manusia untuk memahami data-data tesebut tanpa mampu mengelolah data tersebut. Akibatnya, data yang dikumpulkan dalam repositori data yang besar menjadi "kuburan data" (Han dan Kamber, 2007).

Hal ini melatarbelakangi lahirnya suatu cabang ilmu pengetahuan baru yaitu *data mining. Data mining* adalah untuk mengekstrasikan atau "menambang" pengetahuan dari kumpulan

banyak data. Data mining adalah teknik yang merupakan gabungan metode-metode analisis data secara berkesinambungan dengan algoritma-algoritma untuk memproses data berukuran besar. Data mining merupakan proses menemukan informasi atau pola yang penting dalam basis data berukuran besar dan merupakan kegiatan untuk menemukan informasi atau pengetahuan yang berguna secara otomatis dari data yang jumlahnya besar. *Data mining,* sering juga disebut *knowledge discovery in database* (*KDD),* adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan pola keteraturan, pola hubungan dalam set data berukuran besar(Santosa, 2007). Keluaran dari data mining ini dapat dijadikan untuk memperbaiki pengambilan keputusan di masa depan. Dalam *data mining* data disimpan secara elektronik dan diolah secara otomatis, atau setidaknya disimpan dalam komputer. Data mining adalah tentang menyelesaikan masalah dengan menganalisa data yang telah ada dalam database(Witten dan Frank, 2011).

Berdasarkan tugasnya, *data mining* dikelompokkan menjadi :

1. Deskripsi
   Mencari cara untuk menggambarkan pola dan trend yang terdapat dalam data. Sebagai contoh, seorang pengumpul suara mengungkap bukti bahwa mereka yang diberhentikan dari jabatannya saat ini, akan kurang mendukung dalam pemilihan presiden. Untuk deskripsi ini bisa dilakukan dengan *exploratory data analysis,* yaitu metode grafik untuk menelusuri data dalam mencari pola dan trend.
2. Estimasi
   Estimasi mirip seperti klasifikasi tapi variabel sasaran adalah numerik. Model dibuat menggunakan *record* yang lengkap, juga ada variable targetnya. Kemudian untuk data baru, estimasi nilai variable target dibuat berdasarkan nilai prediktor. Contoh, untuk estimasi tekanan darah pada pasien, variabel prediktornya umur, jenis kelamin, berat badan, dan tingkat sodium darah. Hubungan antara tekanan darah, dan variable prediktor pada data training akan menghasilkan model kemudian diaplikasikan pada data baru. Untuk melakukan estimasi bisa digunakan *neural network* atau metode *statistic* seperti *point estimation* dan *confidence interval estimations, simple linear regression* dan *correlation,* dan *multiple regression*[9].
3. Prediksi
   Prediksi mirip seperti klasifikasi dan estimasi, tapi hasilnya untuk memprediksi masa depan. Contoh, memprediksi harga barang tiga bulan mendatang, memprediksi presentasi kenaikan angka kematian karena kecelakaan tahun mendatang jika kecepatan berkendara dinaikkan. Metode dan teknik untuk klasifikasi dan estimasi, jika cocok, bisa juga digunakan untuk prediksi, termasuk metode statistik. Algoritma untuk prediksi antara lain *regression tree* dan model *tree*.
4. Klasifikasi
   Dalam klasifikasi, sasarannya adalah variabel kategori, misalkan atribut penghasilan, yang bisa dikategorikan menjadi tiga kelas atau kategori yaitu, tinggi, sedang, dan rendah. Model data mining membaca sejumlah besar *record* tiap *record* berisi informasi pada variabel target. Contoh, dari sebuah data set misalkan mau mengklasifikasikan penghasilan seseorang yang datanya tidak terdapat pada dataset, berdasarkan karakteristik yang berhubungan dengan orang itu seperti, umur, jenis kelamin, dan pekerjaan. Tugas klasifikasi ini cocok untuk metode dan teknik *data mining.* Algoritma akan mengolah dengan cara membaca data set yang berisi variabel *predictor* dan variabel taget yang telah diklasifikasikan, yaitu penghasilan. Di sini algoritma (*software)* "mempelajari" kombinasi variabel mana yang berhubungan dengan penghasilan yang mana. Data ini disebut *training set.* Kemudian algoritma akan melihat ke data baru yang belum termasuk klasifikasi manapun. Berdasarkan klasifikasi pada data set kemudian algoritma akan memasukkan data baru tersebut ke dalam klasifikasi yang mana. Misalkan seorang professor wanita berusia 63 tahun bisa jadi diklasifikasikan ke dalam kelas penghasilan tinggi. Algoritma klasifikasi yang banyak digunakan secara luas untuk klasifikasi antara lain *decision tree, bayesian classifier,* dan *neural network*.
5. *Clustering*
   Clustering mengacu pada pengelompokkan *record-record,* observasi, atau kasus-kasus ke dalam kelas-kelas dari objek yang mirip. Pada *clustering* tidak ada variabel sasaran. Sebuah *cluster* adalah koleksi record yang mirip satu sama lain, dan tidak mirip dengan *record* pada *cluster.* Tidak seperti klasifikasi, pada clustering tidak ada variabel target. *Clustering* tidak menglasifikasi atau mengestimasi atau memprediksi tetapi mencari untuk mensegmentasi seluruh data set ke *subgroup* yang *relative* sejenis atau *cluster,* dimana kemiripan record di dalam *cluster* dimaksimalkan dan kemiripan dengan *record* di luar *cluster* diminimalkan.

Contoh *clustering,* untuk akunting dengan tujuan audit untuk mensegmentasi *financial behaviour* ke dalam kategori ramah dan curiga, sebagai alat reduksi dimensi ketika data set memiliki ratusan atribut, untuk *clustering* ekspresi gen, dimana kuantitas gen bisa terlihat mempunyai *behavior* yang mirip. Algoritma untuk *clustering* antara lain, *hierarchical agglomerative clustering, Bayesian clustering, self – organizing feature maps, growing hierarchical self - organizing maps*.

6. Asosiasi

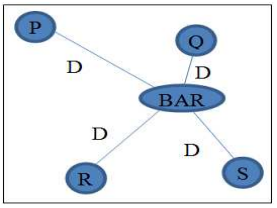
Tugas asosiasi untuk data mining adalah kegiatan untuk mencari atribut yang "go *together".* Dalam dunia bisnis, asosiasi dikenal sebagai *affinity analysis* atau *market basket analysis,* tugas asosiasi adalah membuka *rules* untuk pengukuran hubungan antara dua atribut atau lebih. Contoh asosiasi, prediksi degradasi dalam jaringan komunikasi, menemukan barang apa di supermarket yang dibeli bersama dengan barang lain yang tidak pernah dibeli bersama, menemukan proporsi kasus dimana obat baru akan memperlihatkan efek samping yang berbahaya. Untuk menemukan *association rules,* bisa dilakukan dengan algoritma a priori dan algoritma GRI (*Generalized Rule Induction)I*.

C. Algoritma K-Nearest Neighbor

K-Nearest Neighbor (KNN) termasuk kelompok instance-based learning atau algoritma supervised learning (Kusrini, 2009). Perbedaan antara *supervised learning* dengan *unsupervised learning* adalah pada *supervised learning* bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada dengan data yang baru. Sedangkan pada *unsupervised learning*, data belum memiliki pola apapun, dan tujuan *unsupervised learning* untuk menemukan pola dalam sebuah data.

Algoritma K-Nearest Neighbor (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. K-NN merupakan salah satu metode pengklasifikasian data berdasarkan similaritas dengan label data.

Algoritma ini juga merupakan salah satu teknik lazy learning. kNN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing. Contoh kasus, misal diinginkan untuk mencari solusi terhadap masalah seorang pasien baru dengan menggunakan solusi dari pasien lama. Untuk mencari solusi dari pasien baru tersebut digunakan kedekatan dengan kasus pasien lama, solusi dari kasus lama yang memiliki kedekatan dengan kasus baru digunakan sebagai solusinya. Terdapat pasien baru dan 4 pasien lama, yaitu P, Q, R, dan S (Gambar 1). Ketika ada pasien baru maka yang diambil solusi adalah solusi dari kasus pasien lama yang memikili kedekatan terbesar.



Sumber : (Larose, 2009)

Gambar 1. ilustrasi kasus algoritma KNN

Misal D1 adalah jarak antara pasien baru dengan pasien P, D2 adalah jarak antara pasien baru dengan pasien Q, D3 adalah jarak antara pasien baru dengan pasien R, D4 adalah jarak antara pasien baru dengan pasien S. Dari ilustrasi gambar terlihat bahwa D2 yang paling terdekat dengan kasus baru. Dengan demikian maka solusi dari kasus pasien Q yang akan digunakan sebagai solusi dari pasien baru tersebut.

Ada banyak cara untuk mengukur jarak kedekatan antara data baru dengan data lama (data *training*), diantaranya e*uclidean distance* dan *manhattan distance* (*city block distance*), yang paling sering digunakan adalah *euclidean distance*(Larose, 2005), yaitu:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2} \quad \text{..... (1)}$$

Dimana a = a_1,a_2, …, a_n, dan b = b_1, b_2, …, b_n mewakili n nilai atribut dari dua *record*. Untuk atribut dengan nilai kategori, pengukuran dengan *euclidean distance* tidak cocok. Sebagai penggantinya, digunakan fungsi sebagai berikut (Larose, 2005):

different(a_i,b_i) { 0 jika $a_i = b_i$ = 1 selainnya

Dimana a_i dan b_i adalah nilai kategori. Jika nilai atribut antara dua *record* yang dibandingkan sama maka nilai jaraknya 0, artinya mirip, sebaliknya, jika berbeda maka nilai kedekatannya 1, artinya tidak mirip sama sekali. Misalkan atribut warna dengan nilai merah dan merah, maka nilai kedekatannya 0, jika merah dan biru maka nilai kedekatannya 1.

Untuk mengukur jarak dari atribut yang mempunyai nilai besar, seperti atribut pendapatan, maka dilakukan normalisasi. Normalisasi bisa dilakukan dengan *min-max*

normalization atau *Z-score standardization*[9]. Jika data *training* terdiri dari atribut campuran antara numerik dan kategori, lebih baik gunakan *min-max normalization*.

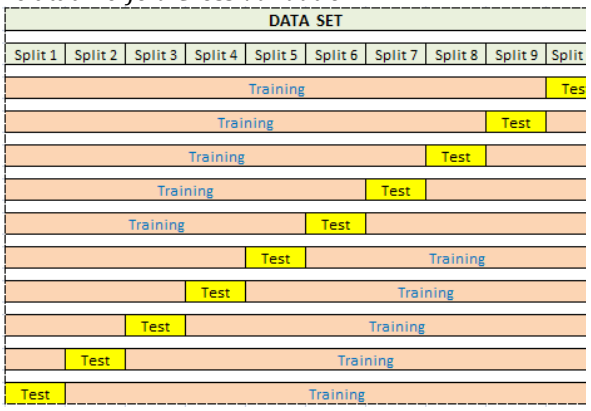
Untuk menghitung kemiripan kasus, digunakan rumus(Kusrini, 2009):

$$\text{Similarity (p, q)} = \frac{\sum_{i=1}^{n} f(pi, qi)\ X\ wi}{wi} \quad \text{..... (2)}$$

Keterangan :
P = Kasus baru
q = Kasus yang ada dalam penyimpanan n
= Jumlah atribut dalam tiap kasus
i = Atribut individu antara 1 sampai dengan n
f = Fungsi *similarity* atribut i antara kasus p dan kasus q
w = Bobot yang diberikan pada atribut ke-i

D. Pengujian K-Fold Cross Validation

Cross Validation adalah teknik validasi dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi(Han dan Kamber, 2007). Dengan menggunakan *cross validation* akan dilakukan percobaan sebanyak k. Data yang digunakan dalam percobaan ini adalah data *training* untuk mencari nilai *error rate* secara keseluruhan. Secara umum pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi. Dalam penelitian ini nilai k yang digunakan berjumlah 10 atau 10-*fold Cross Validation.*



Sumber : (Han dan Kamber, 2007)

Gambar 2. Ilustrasi 10 *Fold Cross Validation*

Pada gambar 4 terlihat bahwa tiap percobaan akan menggunakan satu data *testing* dan k-1 bagian akan menjadi data *training,* kemudian data *testing* itu akan ditukar dengan satu buah data *training* sehingga untuk tiap percobaan akan didapatkan data *testing* yang berbeda-beda.

E. Confusion Matrix

Confusion matrix memberikan keputusan yang diperoleh dalam *traning* dan *testing*, *confusion matrix* memberikan penilaian *performance* klasifikasi berdasarkan objek dengan benar atau salah (Gorunescu, 2011).

Confusion matrix berisi informasi aktual (*actual*) dan prediksi (*predicted)* pada sistem klasifikasi. Berikut tabel penjelasan tentang conusion matrix.

Tabel 1. Confusion Matrix

Classification	Predicted Class		
Observed Class		Class = Yes	Class = No
	Class = Yes	A (True Positif- tp)	B (False negatif- fn)
	Class = No	C (False positif- fp)	D (*true negative*– tn)

Keterangan:
True Positive (tp) = proporsi positif dalam data set yang diklasifikasikan positif
True Negative (tn) = proporsi negative dalam data set yang diklasifikasikan negative
False Positive (fp) = proporsi negatif dalam data set yang diklasifikasikan potitif
FalseNegative(fn) = proporsi negative dalam data set yang diklasifikasikan negatif
Berikut adalah persamaan model *confusion matrix:*

a. Nilai akurasi (acc) adalah proporsi jumlah prediksi yang benar.

Dapat dihitung dengan menggunakan persamaan:

$$acc = \frac{tp+tn}{tp+tn+fp+fn} \quad \text{..... (3)}$$

b. Sensitivity digunakan untuk membandingkan proporsi tp terhadap tupel yang positif, yang dihitung dengan menggunakan persamaan:

$$Sensitivity = \frac{tp}{tp+fn} \quad \text{..... (4)}$$

c. Specificity digunakan untuk membandingan proporsi tn terhadap tupel yang negatif, yang dihitung dengan menggunakan persamaan:

$$Specificity = \frac{tn}{tn+fp} \quad \text{..... (5)}$$

d. PPV (*positive predictive value)* adalan proporsi kasus dengan hasil diagnosa positif, yang dihitung dengan menggunakan persamaan:

$$PPV = \frac{tp}{tp+fp} \quad \text{..... (6)}$$

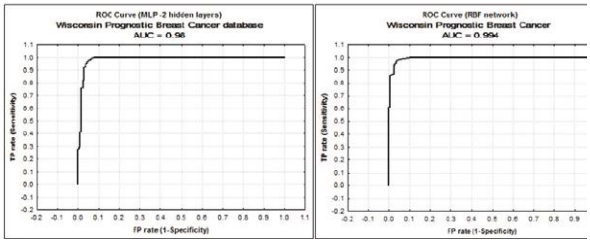
e. NPV (*negative predictive value)* adalah proporsi kasus dengan hasil diagnosa negatif, yang dihitung dengan menggunakan persamaan:

$$NPV = \frac{tn}{tn+fn} \quad \text{..... (7)}$$

F. Curve ROC

Curve ROC (*Receiver Operating Characteristic)* adalah cara lain untuk mengevaluasi akurasi dari klasifikasi secara visual (Vircellis, 2009). Sebuah grafik ROC adalah

plot dua dimensi dengan proporsi positif salah (fp) pada sumbu *X* dan proporsi positif benar (tp) pada sumbu *Y*. Titik (0,1) merupakan klasifikasi yang sempurna terhadap semua kasus positif dan kasus negatif. Nilai positif salah adalah tidak ada (fp = 0) dan nilai positif benar adalah tinggi (tp = 1). Titik (0,0) adalah klasifikasi yang memprediksi setiap kasus menjadi negatif {-1}, dan titik (1,1) adalah klasifikasi yang memprediksi setiap kasus menjadi positif {1}. Grafik ROC menggambarkan *trade-off* antara manfaat ('*true positives*') dan biaya ('*false positives*'). Berikut tampilan dua jenis kurva ROC (*discrete* dan *continous*).



Sumber : (Gorunescu, 2011)

Gambar 3. Grafik ROC (*discrete* dan *continous*)

Pada Gambar 2.4 garis diagonal membagi ruang ROC, yaitu:

1. (a) poin diatas garis diagonal merupakan hasil klasifikasi yang baik.
2. (b) point dibawah garis diagonal merupakan hasil klasifikasi yang buruk.

Dapat disimpulkan bahwa, satu point pada kurva ROC adalah lebih baik dari pada yang lainnya jika arah garis melintang dari kiri bawah ke kanan atas didalam grafik. Tingkat akurasi dapat di diagnosa sebagai berikut[10]:
Akurasi 0.90 – 1.00 = *Excellent classification*
Akurasi 0.80 – 0.90 = *Good classification*
Akurasi 0.70 – 0.80 = *Fair classification*
Akurasi 0.60 – 0.70 = *Poor classification*
Akurasi 0.50 – 0.60 = *Failure*

Metode penelitian

Penelitian adalah mencari melalui proses yang metodis untuk menambahkan pengetahuan itu sendiri dan dengan yang lainnya, oleh penemuan fakta dan wawasan tidak biasa. Pengertian lainnya, penelitian adalah sebuah kegiatan yang bertujuan untuk membuat kontribusi orisinal terhadap ilmu pengetahuan(Dawson, 2009).

Penelitian ini adalah penelitian eksperimen dengan metode penelitian sebagai berikut

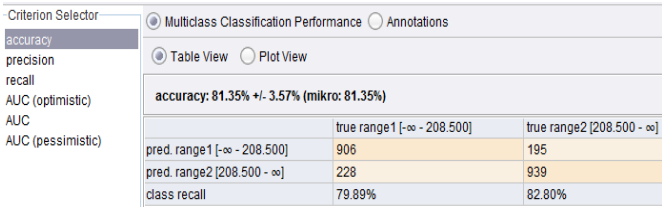
1. Pengumpulan data
   Pada pengumpulan data dijelaskan tentang bagaimana dan darimana data dalam penelitian ini didapatkan, ada dua tipe dalam pengumpulan data, yaitu pengumpulan data primer dan pengumpulan data sekunder. Data primer adalah data yang dikumpulkan pertama kali untuk melihat apa yang sesungguhnya terjadi. Data sekunder adalah data yang sebelumnya pernah dibuat oleh seseorang baik di terbitkan atau tidak. Dalam pengumpulan data primer dalam penelitian ini menggunakan metode observasi dan interview, dengan menggunakan data-data yang berhubungan dengan pemilu ditahun 2009. 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 yang digunakan 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. Sedangkan varaibel tujuannya yaitu hasil pemilu.
2. Pengolahan awal data
   Jumlah data awal yang diperoleh dari pengumpulan data yaitu sebanyak 2.268 data, namun tidak semua data dapat digunakan dan tidak semua atribut digunakan karena harus melalui beberapa tahap pengolahan awal data (*preparation data*). Untuk mendapatkan data yang berkualitas, beberapa teknik yang dilakukan yaitu(Vercelis, 2009): data validation, data integration and transformation dan data size reduction and discritization. Sehingga diperoleh atribut antara lain, jenis kelamin, no.urut parpol, suara sah partai, jumlah perolehan kursi, daerah pemilihan, nomor urut caleg dan suara sah caleg.
3. Model yang diusulkan
   Model yang diusulkan pada penelitian ini berdasarkan *state of the art* tentang prediksi hasil pemilihan umum adalah dengan menerapkan *K-Nearest Neighbor* untuk memprediksi hasil pemilu Legisltif DKI Jakarta, yang terlihat pada Gambar dibawah ini
4. Eksperiment dan pengujian model
   Pada penelitian ini, penulis akan menerapkan algoritma *K-Nearest Neighbor* untuk memprediksi seberapa akurat data yang ada.
5. Evaluasi dan validasi hasil
   Setelah ditemukan nilai akurasi yang paling ideal dari parameter di atas akan terbentuk struktur algoritma yang ideal untuk pemecahan masalah tersebut. Model yang diusulkan pada penelitian tentang prediksi hasil pemilihan umum adalah dengan menerapkan *algoritma K-Nearest Neighbor*.

HASIL DAN PEMBAHASAN

A. Metode K-Nearest Neighbor

Algoritma *K-Neraerst Neighbor* adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.

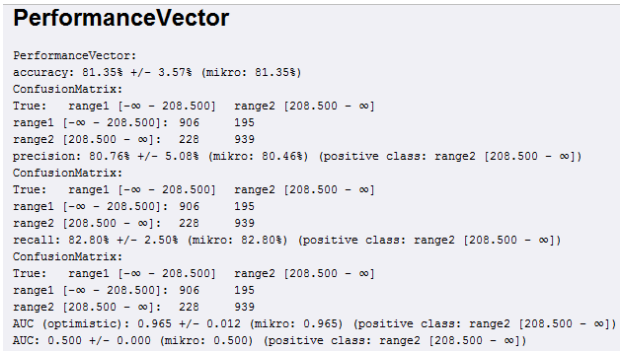
Hasil eksperiment yang penulis lakukan, data sebanyak 2.268 data yang penulis analisis terlihat bahwa nilai akurasi sebesar 81.35 % seperti gambar yang terlihat di bawah ini.



Sumber : (Data Penelitian, 2015)

Gambar 4. Nilai Akurasi

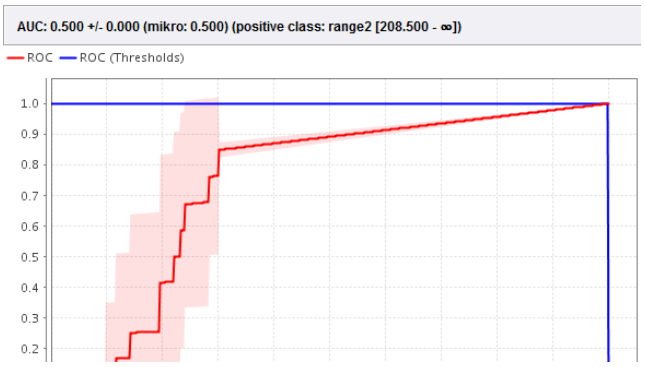
Informasi yang dihasilkan dari pengolahan data tersebut dapat diperoleh gambaran data seperti dua gambar di atas yaitu nilai accuracy dan kurva auc. Disamping informasi gambar yang terlihat seperti di atas, penulis juga menyertakan informasi yang bisa di lihat oleh analis data yaitu performance vector dimana bedanya dengan table accuracy adalah jika table accuracy informasi yang di hasilkan dalam bentuk tabel. Namun jika ingin melihat hasil analisis dalam bentuk teks bisa dilihat dalam bentuk teks yaitu performance vector. Seperti yang terlihat di gambar di bawah ini.



Sumber : (Data Penelitian, 2015)

Gambar 5. Performance Vector

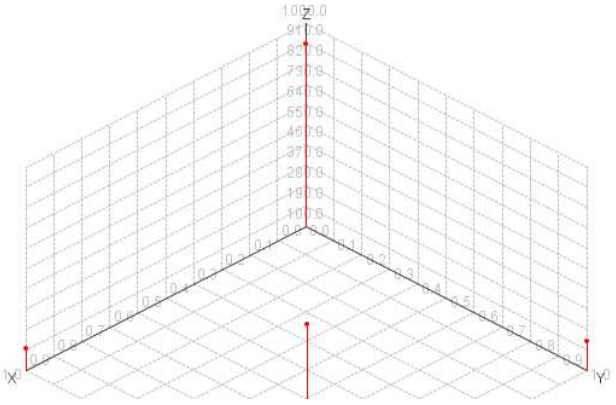
Sedangkan nilai kurva AUC yang terbentuk dari data tersebut adalah 0.500 dengan menggunakan data yang sama seperti gambar yang terlihat dibawah ini.



Sumber : (Data Penelitian, 2015)

Gambar 6. Kurva AUC

Disamping penggambaran Curva AUC seperti di atas, penulis juga menyajikan hasil kurva AUC menggunakan model plot view seperti di bawah ini.



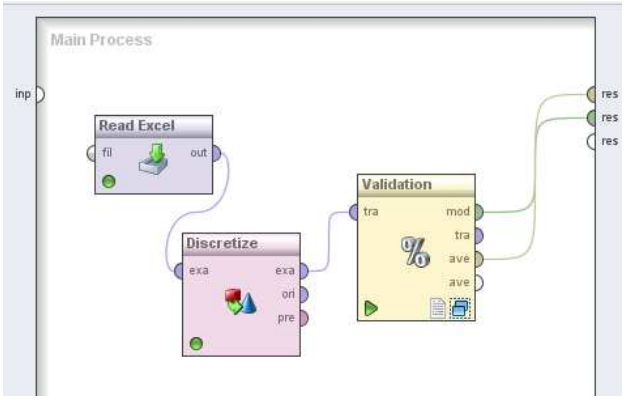
Sumber : (Data Penelitian, 2015)

Gambar 7. Confusion Matrix
Gambar Confusion Matrix Plot View

B. Analisa Evaluasi dan Validasi Model

Hasil dari pengujian model yang dilakukan adalah memprediksi hasil pemilu legislatif DKI Jakarta 2009 dengan menggunakan algoritma *K-Nearest Neighbor* untuk menentukan nilai *accuracy* dan *AUC.*

Dalam menentukan nilai tingkat keakurasian dalam model *K-Nearest Neighbor,* metode pengujian yang dilakukan menggunakan *cross validation* dengan desain modelnya sebagai berikut.



Sumber : (Data Penelitian, 2015)

Gambar 8. Pengujian cross validation

Dari hasil pengujian diatas, baik evaluasi menggunakan *counfusion matrix* maupun ROC *curve* terbukti bahwa hasil pengujian algoritma *K-Nearest Neighbor*.

Nilai akurasi untuk model algoritma *K-Nearest Neighbor* sebesar 81.35 %. Sedangkan evaluasi menggunakan ROC *curve* sehingga menghasilkan nilai AUC (*Area Under Curve)* untuk model algoritma *K-Nearest Neighbor* mengasilkan nilai 0.500 dengan nilai diagnosa *Excellent Classification.*

KESIMPULAN

Berdasarkan hasil eksperiment yang dilakukan dari hasil analisis optimasi model algoritma *K-Nearest Neighbor* sebesar 81.35 % dan nilai AUC sebesar 0.500 dengan tingkat diagnosa *Excellent Classification*. Data ini nantinya bisa digunakan oleh pihak yang membutuhkan informasi terkait pemilihan umum khususnya pemilihan umum di tingkat Daerah I atau Daerah II.

REFERENSI

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

Gorunescu, F. , 2011, *Data Mining Concept Model Technique.* India: Springer.

Han, J., & Kamber, M. , 2007, *Data Mining Concepts and Technique.* Morgan Kaufmann publisher.

Han, J., dan Kamber, M. (2007). *Data Mining Concepts and Technique*. Morgan Kaufmann publisher

Kusrini, & Luthfi, E. T. , 2009, *Algoritma Data mining.* Yogyakarta: Andi.

Larose, D. T. , 2005, *Discovering Knowledge in Data.* Canada: Wiley Interscience.

Nagadevara, & Vishnuprasad. , 2005, Building Predictive models for election result in india an application of classification trees and neural network. *Journal of Academy of Business and Economics Volume 5* .

Rigdon, S. E., Jacobson, S. H., Sewell, E. C., & Rigdon, C. J. , 2009, A Bayesian Prediction Model For the United State Presidential Election. *American Politics Research volume.37* , 700-724.

Santoso, T. (2004). Pelanggaran pemilu 2004 dan penanganannya. Jurnal demokrasi dan Ham , 9-29.

Sardini, N. H. , 2011, *Restorasi penyelenggaraan pemilu di Indonesia.* Yogyakarta: Fajar Media Press.

Undang-Undang RI No.10 , 2008.

Vercellis, C. , 2009, *Business Intelligence : Data Mining and Optimization for Decision Making.* John Wiley & Sons, Ltd.

Witten, I. H., Frank, E., & Hall, M. A. 2011. *Data Mining: Practical Machine Learning and Tools*. Burlington : Morgan Kaufmann Publisher

BIODATA PENULIS

Mohammad Badrul, M.Kom
Penulis adalah Dosen Tetap di STMIK Nusa Mandiri Jakarta. Penulis Kelahiran di Bangkalan 01 Januari 1984. Penulis menyelesaikan Program Studi Strata 1 (S1) di Kampus STMIK Nusa Mandiri Prodi Sistem Informasi dengan gelar S.Kom pada tahun 2009 dan menyelesaikan progarm Srata 2 (S2) di Kampus yang sama dengan Prodi ilmu Komputer dengan gelar M.Kom pada tahun 2012. Selain mengajar, Penulis juga aktif dalam membimbing mahasiswa yang sedang melakukan penelitian khususnya di tingkat Strata 1 dan penulis juga terlibat dalam tim konsorsium di Jurusan Teknik Informatika STMIK Nusa Mandiri untuk penyusunan bahan ajar. Saat ini penulis memiliki Jabatan Fungsional Asisten ahli di kampus STMIK Nusa Mandiri Jakarta. Penulis tertarik dalam bidang kelimuan Data mining, Jaringan komputer, Operating sistem khusunya open source, Database, Software engineering dan Research Metode.