PREDIKSI HASIL PEMILU LEGISLATIF DKI JAKARTA MENGGUNAKAN NAÏVE BAYES DENGAN ALGORITMA GENETIKA SEBAGAI FITUR SELEKSI

Diana Tri Wahyuni[1], T.Sutojo[2], Ardytha Luthfiarta[3]
Jurusan Teknik Informatika FIK UDINUS
Jl. Nakula 1 No. 5-11. Semarang 50131 INDONESIA
[1]dianatriwahyuni99@gmail.com
[2]tsutojo@dosen.dinus.ac.id
[3]ardyt.luthfi@gmail.com

Intisari— **Pemilu merupakan sebuah tonggak untuk menentukan pemimpin dari sebuah negara demokrasi. Hasil pemilu berdampak pada berbagai sektor. Maka dibutuhkan sistem perhitungan cepat atau prediksi untuk memprediksikan hasil pemilu. Saat ini terdapat sebuah sistem perhitungan cepat yang sering disebut *quickcount* untuk menentukan hasil pemilu secara statistik. Akan tetapi untuk mencapai akurasi yang tinggi, harus disertai perhitungan-perhitungan statistik dan pengambilan sampling beserta jumlah sampel yang tepat. Hal ini akan memakan waktu dan biaya. Atas dasar permasalahan tersebut, dilakukan penelitian untuk menerapkan data mining khususnya naïve bayes dalam sistem prediksi hasil pemilu legislatif DKI Jakarta dan menggunakan algoritma genetika sebagai *feature selection*. *Naïve bayes* merupakan algoritma klasifikasi data mining yang menganggap masing-masing atribut tidak saling berhubungan. Maka dari itu digunakan algoritma genetika untuk membantu naïve bayes dalam menentukan atribut-atribut yang harus digunakan sehingga dapat meningkatkan akurasi. Prediksi hasil pemilu legislatif DKI Jakarta menggunakan naïve bayes memiliki akurasi 92,28% dan nilai AUC 0,981. Sedangkan prediksi menggunakan naïve bayes dan AG sebagai fitur seleksi memiliki akurasi 97,84% dan nilai AUC 0,994. jadi penggunaan AG sebagai fitur seleksi dapat meningkatkan akurasi prediksi hasil pemilu legislatif DKI Jakarta.**

Kata kunci— data mining, naïve bayes, prediksi pemilu legislatif DKI Jakarta, algoritma genetika, fitur seleksi.

Abstract— **Election is a milestone to determine the leader of a democratic country . The election results have an impact on various sectors . So we need a quick calculation or prediction system to predict election results . Currently there is a fast computation system is often called quickcount to statistically determine the outcome of the election . However, to achieve high accuracy , must be accompanied by statistical calculations and sampling along with the appropriate number of samples . This will take time and cost .**
On the basis of these issues , conducted research to apply data mining systems , especially naïve Bayes prediction in Jakarta legislative election results and uses a genetic algorithm as a feature selection. Naïve Bayes is a classification of data mining algorithms that consider each attribute are not interconnected . Therefore genetic algorithm is used to assist in determining the naïve Bayes attributes that should be used so as to improve the accuracy . Prediction results of the legislative elections of Jakarta using naïve Bayes have the accuracy 92.28 % and the AUC value 0.981 . While predictions using naïve Bayes and AG as feature selection have the accuracy 97.84 % and the AUC value 0.994 . so the use of AG as a feature selection can improve the accuracy prediction of Jakarta legislative election result.

Keywords— data mining, naïve bayes, legislative election prediction of DKI Jakarta

I. PENDAHULUAN

1.1 Latar Belakang

Pemilu merupakan sebuah tonggak untuk menentukan pemimpin dari sebuah negara demokrasi [1]. Menurut undang-undang ri no.10 tahun 2008 pemilu merupakan sarana untuk melaksanakan kedaulatan rakyat di negara kesatuan republik indonesia yang berlandaskan pancasila dan uud 1945. Pemilu adalah peristiwa politik yang penting bagi terwujudnya sistem politik yang demokratis.

menurut undang-undang ri no.10 tahun 2008, tujuan pemilu adalah untuk menentukan anggota dpr, dprd provinsi, dan dprd kabupaten/kota yang dilakukan dengan sistem proposional terbuka. dengan besarnya jumlah partai dan sistem pemilu langsung, pemilu legislatif membuat rakyat indonesia berpeluang besar untuk menjadi anggota legislatif. pemilu legislatif 2009 diikuti partai nasional dan partai lokal sebanyak 44 partai. pemilu legislatif dki jakarta tahun 2009 terdapat 2268 calon anggota dprd dari 44 partai yang memperebutkan 94 kursi anggota dpr dki jakarta.

hasil pemilu legislatif dapat berdampak pada berbagai aspek. salah satunya adalah dari segi ekonomi. pada penelitian dr. adler haymans manurung dan cahyanti ira k (2006), "sektor-sektor di bursa efek jakarta memberikan reaksi terhadap peristiwa pengumuman hasil pemilu legislatif. abnormal return yang signifikan negatif yang terjadi pada hari ke-3 sebelum pengumuman hasil pemilu legislatif merupakan reaksi atas kekecewaan investor karena komisi pemilihan umum gagal untuk memenuhi jadwalnya untuk mengumumkan hasil pemilu pada tanggal 28 april 2004. sementara itu sentimen negatif pasar masih tetap terjadi hingga hari ke-3 setelah pengumuman hasil pemilu legislatif, yang masih memiliki abnormal return yang negatif signifikan. namun pasar memberikan reaksi yang berlebih (overreaction) terhadap informasi sehingga return saham rebound pada hari-hari berikutnya, hal ini dibuktikan dari nilai abnormal return saham yang positif pada hari-hari berikutnya. informasi yang beredar pada event date ternyata memiliki nilai informasi pada sektor-sektor: industri dasar kimia; aneka industri; industri barang konsumsi; konstruksi, properti dan real estat; keuangan; serta perdagangan, jasa dan investasi"[14].

quick count merupakan salah satu metode yang digunakan untuk melakukan prediksi hasil pemilu dengan menerapkan ilmu statistik. quick count dilakukan dengan menggunakan metode-metode penelitian yang benar, sahih, beretika, terbuka untuk diperiksa akuntabilitasnya, netral dalam pengertian mengedepankan kebenaran nilai-nilai ilmiah. quick count ini merupakan kegiatan pengambilan sampling biasa, sama seperti survey yang sering dilakukan untuk mengkaji objek studi tertentu, perbedaan hanya pada unit terkecil yang diambil dalam sampel. jika survey unit terkecil adalah desa/kelurahan sedangkan quick count ini adalah tps. alasan waktu dan biaya menjadikan proses pengambilan sampling sering dilakukan baik dalam survey maupun quick count[16].

pada penelitian ini, penulis akan melakukan prediksi hasil pemilu legislatif menggunakan teknik data mining yaitu naïve bayes dan optimasi seleksi fitur algoritma genetika. teknologi data mining merupakan salah satu alat bantu untuk penggalian data pada basis data berukuran besar dan dengan spesifikasi tingkat kerumitan yang telah banyak digunakan pada banyak domain aplikasi seperti perbankan maupun bidang telekomunikasi [12]. naïve bayes merupakan algoritma klasifikasi yang efektif (mendapatkan hasil yang akurat) dan efisien (proses penalaran dilakukan memanfaatkan input yang ada dengan cara yang relatif cepat). algoritma ini bertujuan untuk melakukan klasifikasi data pada klas tertentu [6]. dengan menggunakan data mining, dapat dilakukan prediksi hasil pemilihan umum tanpa harus mengetahui jumlah populasi dari objek yang sedang dikaji dan tidak perlu melakukan proses-proses perhitungan statistik seperti yang dilakukan pada quick count.

penelitian mengenai prediksi hasil pemilu telah dilakukan oleh beberapa peneliti. diantaranya adalah dengan menggunakan metode decision tree, pada tahun 1999 choi dan han memprediksi hasil pemilihan presiden di korea [3]. pada tahun 2005, borisyuk, rallings, dan thrasher memprediksi hasil pemilu dengan metode neural network [2]. pada tahun yang sama, nagadevara dan vishnuprasad, dengan model classification tree dan neural network mereka memprediksi hasil pemilihan umum [4]. moscato, mathieson, mendes dan barreta dengan menggunakan metode decision tree mereka memprediksi hasil pemilihan presiden amerika serikat [5]. mohammad badrul pada tahun 2012 memprediksi hasil pemilu legislatif dki jakarta menggunakan metode neural network berbasis particle swarm optimization[13].

pada penelitian ini, penulis akan memprediksi hasil pemilu legislatif dki jakarta menggunakan algoritma naïve bayes dengan fitur seleksi algoritma genetika. algoritma genetika merupakan salah satu algoritma optimasi seleksi fitur. salah satu proses seleksinya adalah dengan mengambil beberapa individu terbaik. selain itu juga dapat dilakukan dengan proses pengambilan acak proporsional, dengan proporsi setara dengan proporsi kualitasnya. artinya individu yang kualitasnya lebih baik memiliki peluang terpilih lebih besar dan pengambilan dilakukan dengan pemulihan[15]. penerapan algoritma genetika pada penelitian ini diharapkan dapat meningkatkan akurasi prediksi. dataset yang penulis gunakan adalah dataset pemilu legislatif dki jakarta 2009 dengan atribut diantaranya adalah nama partai politik, nama calon legislatif, jenis kelamin, kecamatan, nomor urut partai politik, jumlah perolehan kursi, daerah pemilihan, nomor urut calon legislatif, suara sah partai politik, suara sah calon legislatif[13].

1.2 Rumusan Masalah

Berdasarkan latar belakang yang penulis kemukakan di atas, maka dalam penelitian ini dapat diambil rumusan masalah sebagai berikut **"**bagaimana peningkatan akurasi naïve bayes jika algoritma genetika diterapkan untuk pemilihan atribut yang sesuai dan optimal pada naïve bayes?".

1.3 Batasan Masalah

Batasan masalah untuk penelitian ini dibatasi oleh penulis, yaitu meliputi :

a. Penelitian ini diberlakukan untuk pemilu legislatif DKI Jakarta.
b. Penelitian ini menggunakan data pemilu legislatif DKI Jakarta tahun 2009.
c. Penelitian ini diaplikasikan menggunakan framework rapid miner.
d. Penelitian ini dibatasai pada penerapan algoritma naïve bayes dan optimasi seleksi fitur dengan algoritma genetika untuk prediksi hasil pemilihan Legislatif DKI Jakarta.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah diatas maka tujuan penelitian ini adalah menerapkan algoritma genetika untuk memilih atribut dari dataset untuk meningkatkan akurasi hasil prediksi pemilu legislatif DKI Jakarta menggunakan algoritma naïve bayes.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah :

a. Bagi Penulis, penelitian ini berguna untuk memperluas pengetahuan mengenai data mining khususnya mengenai metode *naïve bayes* beserta penerapannya dan algoritma genetika beserta penerapannya.
b. Bagi masyarakat, penelitian ini dapat memprediksi hasil pemilu legislatif DKI Jakarta, sehingga dapat digunakan oleh para calon legislatif atau partai politik, untuk menentukan strategi memenangkan pemilu legislatif DKI Jakarta. Selain itu juga dapat membantu para pelaku bisnis dan lain sebagainya dalam menentukan kebijakan bisnis dan atau yang lainnya yang memiliki keterkaitan kepentingan dengan hasil pemilu legislatif DKI Jakarta.
c. Hasil penelitian ini juga dapat dimanfaatkan sebagai referensi untuk penelitian selanjutnya mengenai prediksi hasil pemilu legislatif dengan menerapkan algoritma yang berbeda atau dapat melakukan seleksi atribut dengan metode-metode yang ada sehingga dapat meningkatkan keakurasiannya.
d. Memberikan kontribusi keilmuan pada penelitian algoritma naïve bayes berbasis algoritma genetika sebagai strategi untuk memenangkan pemilu legislatif.

II. LANDASAN TEORI

2.1 Pemilu

Pemilihan umum merupakan salah satu tonggak utama dari sebuah demokrasi. Dalam negara demokrasi, pemilihan umum merupakan salah satu tonggak utama untuk memilih pemimpin yang akan mewakili rakyat untuk duduk dipemerintahan mulai dari anggota DPRD tingkat II, DPRD Tingkat I, DPR RI dan DPD.

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[13]. 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. 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. Hasil perolehan suara ditetapkan oleh KPU secara nasional, berikut cara pembagian perhitungan suara setiap periodenya [26]:

1. Pemilu tahun 1955

Penentuan kursi di tiap daerah benar-benar didasarkan pada proporsi jumlah penduduk.

2.. Pemilu tahun1971

Pembagian kursi dibagi habis disetiap daerah pemilihan. Dengan cara ini mampu mengurangi jumlah partai yang meraih kursi dengan sistem kombinasi, kelemahannya menyebabkan suara partai terbuang.

3. Pemilu tahun 1977 – 1997

Model pembagian kursi masih sama dengan pemilu ditahun 1971.

4. Pemilu tahun 1999

Cara pembagian kursi menggunakan sistem proporsional setengah terbuka, artinya jumlah suara yang masuk ke partai dibagi secara proporsional terhadap jumlah pembagi per satu kursi anggota.

5. Pemilu tahun 2004

Pembagian kursi dengan cara menghitung suara sah setiap parpol dari satu daerah pemilihan, lalu menghitung bilangan pembagi pemilih (BPP) yaitu total suara sah satu daerah pemilihan dibagi jumlah kursi yang diperebutkan. Tahap pertama parpol yang jumlah suara sahnya lebih dari angka BPP akan langsung mendapat kursi. Tahap kedua sisa kursi akan diberikan kepada parpol satu persatu berdasarkan urutan parpol dengan sisa suara terbanyak. Perhitungan suara ini diatur pada UU No. 12 tahun 2003.

6. Pemilu tahun 2009

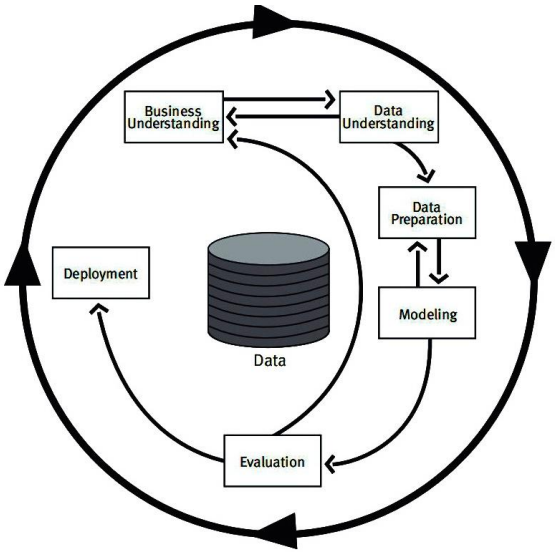
Ditetapkan terlebih dahulu angka bilangan pembagi pemilih (BPP) berdasarkan total suara sah satu daerah dibagi jumlah kursi yang diperebutkan. Penetapan calon terpilih anggota DPR disetiap pemilihan didasarkan atas peringkat suara sah terbanyak pertama, kedua, ketiga, dan seterusnya.

2.2 Data Mining

Secara sederhana *data mining* adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar[8]. *Data mining*, sering juga disebut sebagai *knowledge discovery in database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar [7]. *Data mining* adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam *database*, *data warehouse*, atau penyimpanan informasi lainnya. *Data mining* berkaitan dengan bidang ilmu – ilmu lain, seperti *database system*, *data warehousing*, statistik, *machine learning*, *information retrieval*, dan komputasi tingkat tinggi. Selain itu, *data mining* didukung oleh ilmu lain seperti *neural networ*k, pengenalan pola, *spatial data analysis*, *image database*, *signal processing* [7].

2.3 CRISP-DM

CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) merupakan suatu konsorsium perusahaan yang didirikan oleh Komisi Eropa pada tahun 1996 dan telah ditetapkan sebagai proses standar dalam *data mining* yang dapat diaplikasikan di berbagai sektor industri. Gambar 2.2 menjelaskan tentang siklus hidup pengembangan *data mining* yang telah ditetapkan dalam CRISP-DM.



Gbr. 1 Siklus Hidup CRISP-DM [25]

Berikut ini adalah enam tahap siklus hidup pengembangan *data mining* [25] :

1. *Business Understanding*

Tahap pertama adalah memahami tujuan dan kebutuhan dari sudut pandang bisnis, kemudian menterjemakan pengetahuan ini ke dalam pendefinisian masalah dalam *data mining*. Selanjutnya akan ditentukan rencana dan strategi untuk mencapai tujuan tersebut.

2. *Data Understanding*

Tahap ini dimulai dengan pengumpulan data yang kemudian akan dilanjutkan dengan proses untuk mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi adanya bagian yang menarik dari data yang dapat digunakan untuk hipotesa untuk informasi yang tersembunyi.

3. *Data Preparation*

Tahap ini meliputi semua kegiatan untuk membangun *dataset* akhir (data yang akan diproses pada tahap pemodelan/*modeling*) dari data mentah. Tahap ini dapat diulang beberapa kali. Pada tahap ini juga mencakup pemilihan tabel, *record*, dan atribut-atribut data, termasuk proses pembersihan dan transformasi data untuk kemudian dijadikan masukan dalam tahap pemodelan (*modeling*).

4. *Modeling*

Dalam tahap ini akan dilakukan pemilihan dan penerapan berbagai teknik pemodelan dan beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal. Secara khusus, ada beberapa teknik berbeda yang dapat diterapkan untuk masalah *data mining* yang sama. Di pihak lain ada teknik pemodelan yang membutuhan format data khusus. Sehingga pada tahap ini masih memungkinan kembali ke tahap sebelumnya.

5. *Evaluation*

Pada tahap ini, model sudah terbentuk dan diharapkan memiliki kualitas baik jika dilihat dari sudut pandang analisa data. Pada tahap ini akan dilakukan evaluasi terhadap keefektifan dan kualitas model sebelum digunakan dan menentukan apakah model dapat mencapat tujuan yang ditetapkan pada fase awal (*Business Understanding*). Kunci dari tahap ini adalah menentukan apakah ada masalah bisnis yang belum dipertimbangkan. Di akhir dari tahap ini harus ditentukan penggunaan hasil proses *data mining*.

6. *Deployment*

Pada tahap ini, pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Tahap *deployment* dapat berupa pembuatan laporan sederhana atau mengimplementasikan proses *data mining* yang berulang dalam perusahaan. Dalam banyak kasus, tahap *deployment* melibatkan konsumen, di samping analis data, karena sangat penting bagi konsumen untuk memahami tindakan apa yang harus dilakukan untuk menggunakan model yang telah dibuat.

2.4 Naive Bayes Classifier

Naïve bayes classifier adalah salah satu algoritma dalam teknik data mining yang menerapkan teori Bayes dalam klasifikasi[7]. NBC merupakan pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. NBC terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar.

Teorema Bayes memiliki bentuk umum sebagai berikut.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Keterangan:

X = data dengan class yang belum diketahui

H = hipotesis data X merupakan suatu class spesifik

P(H|X) = probabilitas hipotesis H berdasar kondisi X (posteriori probability)

P(H) = probabilitas hipotesis H (prior probability)

P(X|H) = probabilitas X berdasar kondisi pada hipotesis H

P(X) = probabilitas dari X

Dalam terminologi sederhana, sebuah NBC mengasumsikan bahwa kehadiran (atau ketiadaan) fitur tertentu dari suatu kelas tidak berhubungan dengan kehadiran (atau ketiadaan) fitur lainnya. Sebagai contoh, buah mungkin dianggap apel jika merah, bulat, dan berdiameter sekitar 4 inchi. Bahkan jika fitur ini bergantung satu sama lain atau atas keberadaan fitur lain,. Sebuah NBC menganggap bahwa seluruh sifat-sifat berkontribusi mandiri untuk probabilitas bahwa buah ini adalah apel. Tergantung pada situasi yang tepat dari model probabilitas, NBC dapat dilatih sangat efisien dalam *supervised learning*.

2.5 Algoritma Genetika

Algoritma genetika (AG) diperkenalkan pertama kali oleh John Holland (1975) dari Universitas Michigan, John Holland mengatakan bahwa setiap masalah yang berbentuk adaptasi (alami maupun buatan) dapat diformulasikan ke dalam terminologi genetika [29].

Algoritma genetika merupakan suatu algoritma pencarian berdasarkan pada mekanisme seleksi alam dan genetika alam. Algoritma genetika dimulai dengan sekumpulan solusi awal (individu) yang disebut populasi. Satu hal yang sangat penting adalah bahwa satu individu menyatakan satu solusi. Populasi awal akan berevolusi menjadi populasi baru melalui serangkaian iterasi (generasi). Pada akhir iterasi, algoritma genetika mengembalikan satu anggota populasi yang terbaik sebagai solusi untuk masalah yang dihadapi. Pada setiap iterasi, proses evolusi yang terjadi adalah sebahai berikut[28]:

i. Dua individu dipilih sebagai orang tua (parent) berdasarkan mekanisme tertentu. Kedua parent ini kemudian dikawinkan melalui operator crossover (kawin silang) untuk menghasilkan dua individu anak atau offspring.
ii. Dengan probabilitas tertentu, dua individu anak ini mungkin mengalami perubahan gen melalui operator mutation.
iii. Suatu skema penggantian (replacement scheme) tertentu diterapkan sehingga menghasilkan populasi baru.
iv. Proses ini terus berulang sampai kondisi berhenti (stopping condition) tertentu. Kondisi berhenti bisa berupa jumlah iterasi tertentu, waktu tertentu, atau ketika variansi individu-individu dalam populasi tersebut sudah lebih kecil dari suatu nilai tertentu yang diinginkan.

2.6 Feature Selection

Feature selection adalah sebuah proses yang bisa digunakan pada machine learning dimana sekumpulan dari features yang dimiliki data digunakan untuk pembelajaran algoritma[31]. Subset yang baik memiliki sedikitnya dimensi angka yang paling banyak berkontribusi untuk akurasi dan nantinya akan dibuang sisa dari dimensi yang tidak berkepentingan. Ini merupakan langkah penting dalam tahap preprocessing.

Forward selection dimulai tanpa variabel dan menambahkan mereka satu persatu, pada setiap langkah ditambahkan variable yang menurunkan error paling banyak, sampai semua error dihilangkan[31].

Backward selection dimulai dengan semua variabel dan membuangnya satu persatu, pada setiap langkah membuang variable yang memiliki error paling banyak[31].

2.7 Evaluasi dan Validasi Klasifikasi Data Mining

Untuk melakukan evaluasi pada algoritma naïve bayes dan algoritma naïve bayes dioptimasi dengan algoritma genetika dilakukan beberapa pengujian menggunakan confusion matrix dan kurva ROC (receiver operating charateristic).

2.7.1 Confusion Matrix

Metode *confusion matrix* merepresentasikan hasil evaluasi model dengan menggunakan tabel matriks, jika dataset terdiri dari dua kelas, kelas pertama dianggap positif, dan kelas kedua dianggap negative[24]. Evaluasi menggunakan *confusion matrix* menghasilkan nilai akurasi, presisi, *recall*. Akurasi dalam klasifikasi merupakan presentase ketepatan record data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi[32]. *Precision* atau *confidence* merupakan proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. *Recall* atau *sensitivity* merupakan proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar[33].

Tabel 1. *Confusion Matrix*

Correct Classification	Classified as	
	+	-
+	*True positives*	*False negatives*
-	*False positives*	*True negatives*

True positive (tp) merupakan jumlah *record* positif dalam data set yang diklasifikasikan *positive. True negative* (tn) merupakan jumlah *record negative* dalam data set yang diklasifikasikan *negative. False positive* (fp) merupakan jumlah *record* negatif dalam data set yang diklasifikasikan potitif. *False negative* (fn) merupakan jumlah *record positive* dalam data set yang diklasifikasikan *negative.*

Berikut adalah persamaan model *confusion matrix:*

a. Nilai akurasi (acc) adalah proporsi jumlah prediksi yang benar. Dapat dihitung dengan menggunakan persamaan:

$$akurasi = \frac{tp + tn}{tp + tn + fp + fn}$$

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

$$Sensitivity = \frac{tp}{tp + fn}$$

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

$$Specificity = \frac{tn}{tn + fp}$$

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

$$PPV = \frac{tp}{tp + fp}$$

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

$$NPV = \frac{tn}{tn + fn}$$

2.7.2 Curve ROC

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horizontal dan *true positive* sebagai garis vertical[22].

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[34]:

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

III. METODE PENELITIAN

3.1 Instrumen Penelitian

Sebagai instrument penelitian dibuat program soft computing untuk menguji *proposed method* dan mengukur akurasinya menggunakan rapidminer 5. Sedangkan implementasi naïve bayes untuk prediksi hasil pemilu legislatif DKI Jakarta akan dibangun soft computing menggunakan matlab.

3.2 Metode Pengumpulan Data

Studi kepustakaan digunakan untuk mendapatkan dataset pemilu DKI Jakarta 2009 dengan jumlah data sebanyak 2268 *record*, terdiri dari 11 variabel atau atribut. Variable tersebut ada yang tergolong variabel prediktor atau pemrediksi yaitu variabel yang dijadikan sebagai penentu hasil pemilu, dan variabel tujuan yaitu variabel yang dijadikan sebagai hasil pemilu. 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. Sedangkan variabel tujuannya yaitu hasil pemilu.

Variabel nama partai menunjukkan nama partai politik yang terdaftar sebagai peserta pemilu, dan memiliki tipe data nominal. Variabel nama calon legislatif menunjukkan nama seseorang yang terdaftar sebagai peserta pemilu legislatif DKI jakarta 2009 dari setiap partai, dan bertipe data nominal. Variabel jenis kelamin menunjukkan jenis kelamin dari setiap calon legislatif. Variabel jenis kelamin ini bertipe data nominal. Variabel kota administrasi menunjukkan wilayah kota/kabupaten dari setiap calon legislatif, memiliki tipe data nominal. Variabel no urut parpol menunjukkan nomor urut partai politik yang terdaftar sebagai peserta pemilu legislatif DKI Jakarta 2009. Variabel ini bertipe data ordinal. Variabel suara sah partai menunjukkan jumlah suara sah yang diperoleh setiap partai, bertipe data kontinyu. Variabel jumlah perolehan kursi menunjukkan jumlah kursi yang diperoleh setiap partai politik, bertipe data kontinyu. Variabel daerah pemilihan menunjukkan daerah dimana calon dipilih oleh rakyat. Variabel ini bertipe data ordinal. Variabel no urut caleg menunjukkan nomor urut calon legislatif, bertipe data ordinal. Variabel suara sah menunjukkan jumlah suara sah yang diperoleh setiap calon legislatif. Variabel ini memiliki tipe data kontinyu. Variabel hasil pemilu merupakan variabel target yang menunjukkan hasil dari pemilu yaitu terpilih atau tidak dan bertipe data nominal.

3.3 Teknik Analisis 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 sebagai berikut (Vercellis, 2009):

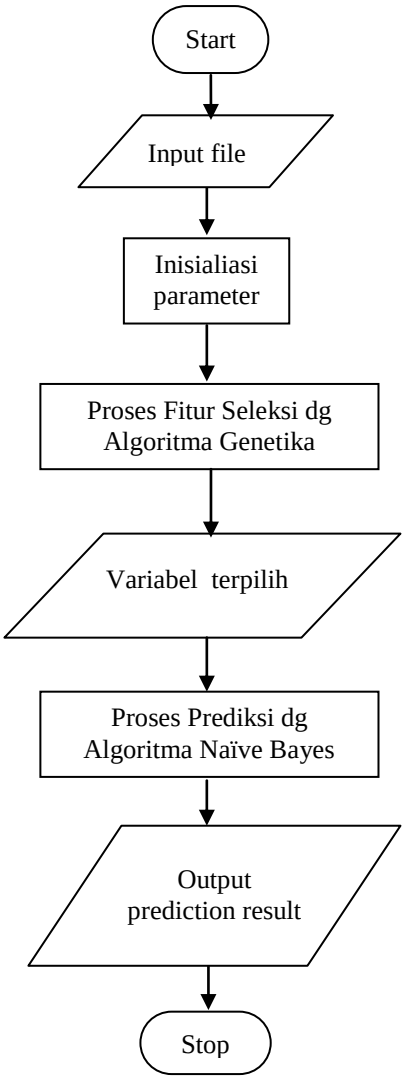
1. Data validation, mengidentifikasikan dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*). Jumlah data awal adalah 2.268 data. Dari keseluruhan data terdapat satu data yang missing, yaitu data ke 2172, pada variable kota administrasi. Sehingga data tersebut dihapus. Data yang dapat digunakan sejauh ini adalah 2267 record.
2. Data *integration and transformation*, meningkatkan akurasi dan efisiensi algoritma. Algoritma naïve bayes dapat memproses data yang bernilai nominal, ordinal, maupun kontinyu. Sehingga nilai-nilai dari setiap atribut yang terdapat pada data set tidak perlu ditransformasikan.
3. *Data size reduction and discritization*, memperoleh data set dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat *informative*. Variabel nomor urut partai dapat dihilangkan. Karena sudah ada variable nama partai politik. Menurut Iberamsjah, seorang pengamat politik Universitas Indonesia, nomor urut partai politik tidak berpengaruh terhadap perolehan suara, karena masyarakat sudah cerdas dan realistis, melihat kinerja parpol yang buruk selama beberapa tahun terakhir. Sehingga pada saat pemilihan, pemilih akan lebih mengingat nama partai politik dibandingkan nomor urut parpol. Dan lagi, pada setiap periode pemilu yang akan datang, nomor urut dari setiap parpol akan berubah. Berbeda dengan nomor urut caleg dan nama caleg, karena banyaknya caleg, tidak mungkin bagi para pemilih untuk mengingat nama-nama caleg yang ada. Pemilih dapat dengan mudah mengingat nomor urut caleg. Sehingga pada waktu pemilihan umum, yang perlu diingat adalah no urut caleg dan dari partai apa. Variabel yang lain tetap diikutkan. Sampai pada tahap terakhir dari persiapan data diperoleh data set dengan jumlah data sebanyak 2.267 data dengan jumlah atribut sebanyak 9 atribut, yaitu nama partai politik, nama caleg, no. urut caleg, jenis kelamin, kota administrasi, daerah pemilihan, suara sah caleg, suara sah partai, jumlah perolehan kursi.

2.4 Metode yang Diusulkan

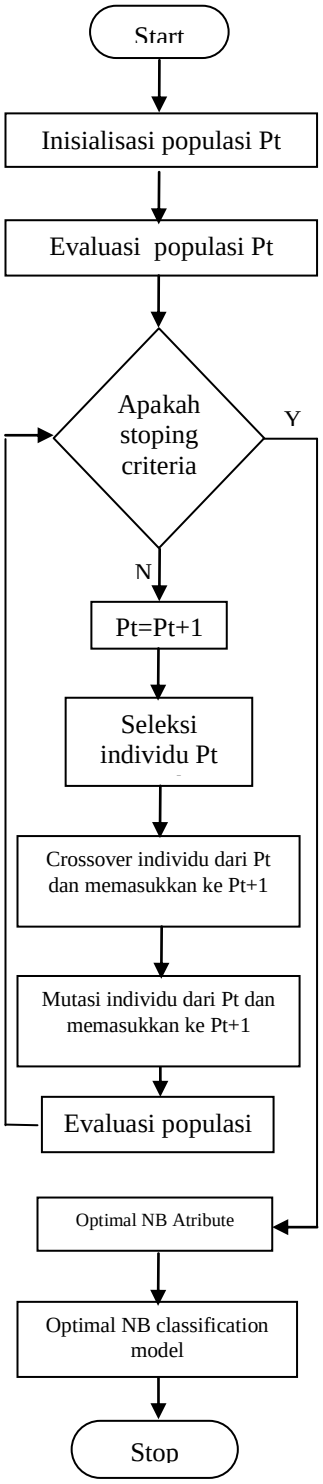
Model yang diusulkan pada penelitian ini adalah dengan menerapkan *naïve bayes* berbasis *Genetic optimization*, yang terlihat pada Gambar 3.3 dan Gambar 3.4. Gambar tersebut menunjukkan proses yang dilakukan dalam tahap modeling untuk menyelesaikan prediksi hasil pemilihan umum dengan menggunakan algoritma *naïve bayes* dengan *Genetic optimization* (algoritma genetika).

Gambar 3.3 Diagram alur utama dari model yang diusulkan. Dimulai dengan memasukkan file data yaitu data set pemilu DKI Jakarta 2009. Kemudian data dengan 9 atribut tersebut akan diseleksi oleh algoritma genetika. Atribut apa saja yang akan digunakan untuk proses prediksi hasil pemilu DKI Jakarta sehingga dapat meningkatkan akurasi prediksi. Sebelum diseleksi, harus ditentukan dahulu parameter-parameter algoritma genetika, seperti *population size, probability crossover, probability mutation.* Setelah terpilih atribut-atribut apa saja yang akan dijadikan sebagai atribut prediktor, kemudian dilakukan training dan testing pada data tersebut menggunakan algoritma naïve bayes. Sehingga naïve bayes dapat melakukan prediksi terhadap hasil pemilu.

Gambar 3.4 menunjukkan diagram alur utama algoritma genetika dan naïve bayes. Langkah pertama dari algoritma genetika adalah inisialisasi populasi Pt. Kromosom-kromosom pada populasi ditentukan nilai gennya. Langkah selanjutnya evaluasi populasi Pt. Kromosom- kromosom diseleksi menggunakan nilai fitness. Kromosom yang memiliki nilai fitness terbesar yang akan terpilih. *Stopping criteria*-nya adalah jumlah maksimal generasi. Jika belum sampai pada generasi maksimal, iterasi akan terus berjalan. Kemudian berdasarkan *probability crossover*, kromosom-kromosom terpilih akan dicrossoverkan. Dan berdasarkan *probability mutation* ditentukan berapa banyak gen dalam kromosom yang akan dimutasi. Setelah mencapai generasi maksimal, akan didapatkan kromosom dengan nilai fitness tertinggi sebagai solusi dari permasalahan seleksi atribut. Kemudian data dengan atribut yang terpilih akan ditraining dan ditesting oleh algoritma naïve bayes, sehingga naïve bayes dapat melakukan prediksi terhadap hasil pemilu DKI Jakarta.



Gambar 3.3 Diagram alur utama



Gambar 3.4 Diagram alur algoritma genetika dan naïve bayes

3.5 Eksperimen dan Pengujian Model

Penelitian ini merupakan penelitian *Experiment*. Dalam penelitian eksperimen digunakan spesifikasi software dan hardware sebagai alat bantu dalam penelitian pada Tabel 3.3:

Software	Hardware
Sistem operasi Windows 7	CPU : intel pentium dual core
Data mining : Rapidminer 5.3	Memory : 2 GB
Matlab R2010a	Hardisk : 250 GB

Untuk mendapatkan variable-variabel yang tepat dan menghasilkan nilai akurasi yang terbesar diperlukan pengaturan untuk parameter-parameter *genetic optimization.* Berikut adalah parameter-parameter yang membutuhkan *adjustment* [19]:

1. Ukuran Populasi *(Pop_Size)*

Populasi adalah kumpulan beberapa individu yang sejenis yang hidup dan saling berinteraksi bersama pada suatu tempat. Jumlah individu dinyatakan sebagai ukuran dari populasi tersebut.

2. P *Crossover (Probability Crossover)*

Pada saat proses genetika berlangsung, nilai dari p crossover digunakan untuk menentukan individu-individu yang akan mengalami crossover.

3. P *Mutation (Probability Mutation)*

Nilai dari p mutasi digunakan untuk menentukan individu yang akan mengalami mutasi, terjadi setelah proses *crossover* dilakukan.

Untuk melihat pengaruh fitur seleksi algoritma genetika terhadap prediksi menggunakan algoritma naïve bayes, penulis akan membandingkan akurasi antara prediksi hasil pemilu legislatif DKI Jakarta menggunakan naïve bayes dan prediksi hasil pemilu legislatif DKI Jakarta menggunakan naïve bayes dengan fitur seleksi algoritma genetika.

3.5.1 Prediksi Hasil Pemilu Legislatif DKI Jakarta menggunakan Naïve Bayes

Naïve bayes akan memprediksi hasil pemilu legislatif DKI Jakarta dengan jumlah atribut prediktor 9, yaitu jenis kelamin, nama parpol, kota administrasi, daerah pemilihan, suara sah caleg, suara sah partai, jumlah perolehan kursi, no urut caleg, nama caleg. Semua atribut akan dihitung peluangnya oleh naïve bayes.

3.5.2 Prediksi Hasil Pemilu Legislatif DKI Jakarta menggunakan Naïve Bayes dengan Fitur Seleksi Algoritma Genetika

Naïve bayes akan melakukan prediksi hasil pemilu legislatif DKI Jakarta dengan atribut prediktor adalah atribut yang telah dipilih oleh algoritma genetika.

Penulis menentukan jumlah minimal fitur yang akan dikombinasi adalah 1 atribut. Sedangkan jumlah maksimal fitur yang akan dikombinasi adalah 9 atribut. Skema seleksi yang digunakan adalah *roullette wheel*. Sedangkan *crossover type*-nya uniform.

Untuk jumlah generasi (*number of generation*), *adjustment* dimulai dari 10 – 100, untuk *Pop size*, *adjustment* dimulai dari 50 - 1000 Untuk P crossover, *adjustment* dimulai dari 0.1 – 1.0 P mutation dari -1.0 – 1.0.

Tabel 3.4 Rencana eksperimen

Num of generation	*Pop Size*	*P Crossover*	*P Mutation*	Jumlah Atribut	Akurasi	AUC
10-100	50-1000	0.1-1.0	-1.0-1.0	?	?	?

Evaluasi akurasi dilakukan oleh confusion matrix dan curve ROC baik prediksi menggunakan naïve bayes maupun naïve bayes dengan fitur seleksi algoritma genetika. Sedangkan pengujian dilakukan dengan 10-fold validation.

3.5.3 Gambaran Umum Program

Setelah mendapatkan atribut prediktor pilihan algoritma genetika, penulis akan mengimplementasikan naïve bayes untuk prediksi hasil pemilu legislatif DKI Jakarta terhadap sebuah program matlab.

Di program tersebut, data yang digunakan naïve bayes untuk training adalah data set pemilu legislatif DKI Jakarta 2009. Di dalam program akan ada inputan nilai atribut-atribut yang telah terpilih oleh algoritma genetika, kemudian naïve bayes akan melakukan perhitungan peluang terhadapnya berdasarkan proses training, hingga menghasilkan output terpilih atau tidak terpilih.

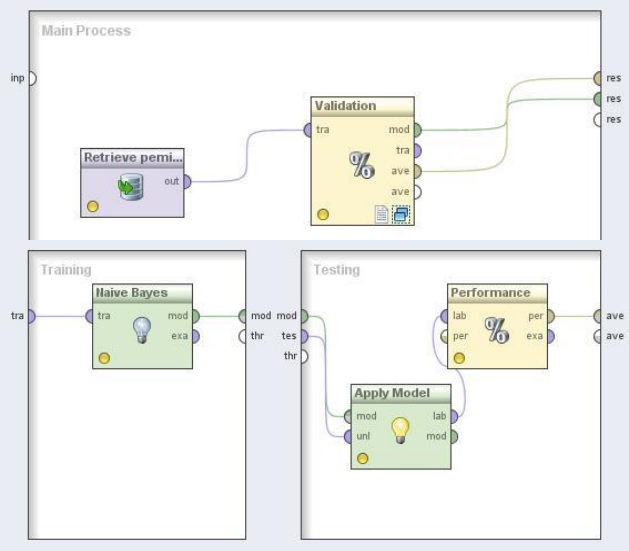
IV. ANALISIS HASIL PENELITIAN DAN PEMBAHASAN

4.1 Eksperimen dan Pengujian Metode

4.1.1 Metode pada Algoritma Naïve Bayes

Algoritma naïve bayes classifier adalah salah satu algoritma dalam teknik data mining yang menerapkan teori Bayes dalam klasifikasi [7]. NBC merupakan pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. NBC terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar[7].

Percobaan pertama adalah menerapkan algoritma naïve bayes untuk memprediksi hasil pemilu DKI Jakarta tanpa menggunakan fitur seleksi. Terdapat 2267 data dan 9 atribut diantaranya adalah nama parpol, nama caleg, no urut caleg, jenis kelamin, kota administrasi, daerah pemilihan, suara sah caleg, suara sah partai, jumlah perolehan kursi. Selain itu juga digunakan *cross validation* untuk melakukan pengujian model.



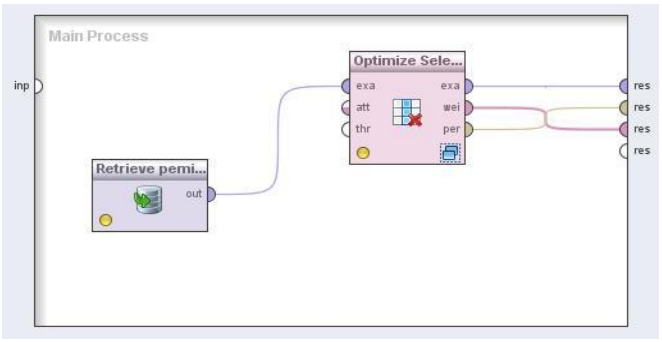
Gambar 4.1 Pemodelan naïve bayes dan *cross validation*

Pada gambar di atas dataset pemilu legislatif DKI Jakarta dihubungkan dengan operator cross validation yang di dalamnya terdapat proses seperti pada gambar 4.1.

Cross validation yang digunakan dalam penelitian ini adalah 10-fold validation. Dataset yang berisi 2267 data dengan 9 atribut akan dipecah menjadi 10 bagian. Dimana setiap bagian akan dibentuk secara random. Prinsip 10-fold validation adalah 1:9, 1 bagian menjadi data testing, data lainnya menjadi data training. Demikian sehingga 10 bagian tersebut berkesempatan menjadi data testing. Setelah dilakukan training dan testing maka dapat diukur akurasinya.

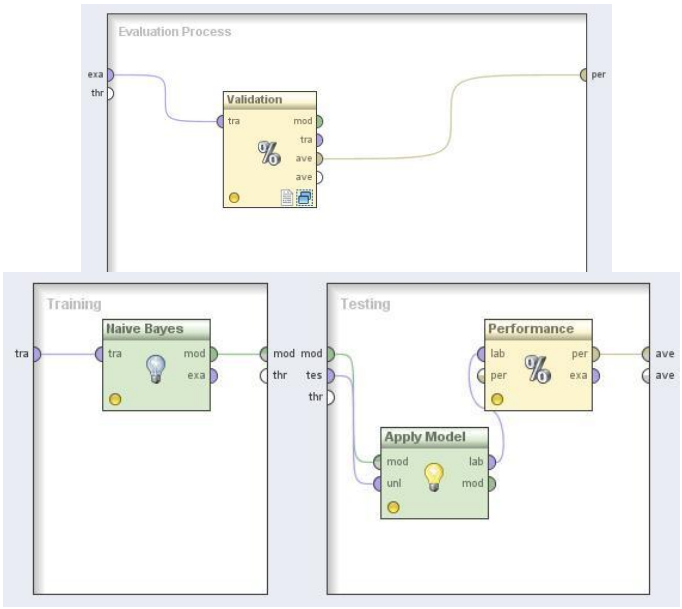
4.1.2 Eksperimen dan Pengujian Metode pada Algoritma Naïve Bayes dengan Fitur Seleksi Algoritma Genetika

Algoritma naïve bayes merupakan algoritma yang tidak memandang keterkaitan antara atribut satu dengan yang lainnya (independensi). Jadi ketika ada data set dengan jumlah atribut ratusan pun, akan dihitung semuanya oleh naïve bayes. Maka dari itu penulis menggunakan algoritma genetika sebagai fitur seleksi, yaitu menentukan atribut-atribut yang relevan sehingga dapat mengoptimalkan akurasi prediksi naïve bayes.



Gambar 4.3 Pemodelan naïve bayes dan algoritma genetika

Dataset pemilu legislatif DKI Jakarta dihubungkan dengan operator *optimize selection (evolutionary)* untuk dilakukan pemilihan atribut-atribut yang relevan dengan proses prediksi hasil pemilu legislatif DKI Jakarta. Di dalam operator *optimize selection (evolutionary)* terdapat proses *cross validation* seperti yang terlihat pada gambar 4.4.



Gambar 4.4 Pengujian model menggunakan *cross validation*

Cross validation membagi data menjadi 10 bagian. Seperti pada proses sebelumnya. Di dalam operator *cross validation* terdapat proses seperti pada gambar 4.5 berikut.

Gambar 4.5 Pemodelan dan pengujian naïve bayes dan AG

Naïve bayes melakukan training terhadap data-data yang telah dibagi oleh *cross validation*. Setelah dilakukan training dan testing dapat dihitung akurasi dari penerapan algoritma genetika dan naïve bayes untuk proses prediksi hasil pemilu legislatif DKI Jakarta.

Batasan jumlah atribut adalah 1-9. Sehingga susunan gen dalam kromosom akan memiliki banyak kombinasi antar 9 atribut. Skema seleksi yang digunakan adalah *roulette wheel*.

Percobaan dimulai dengan melakukan *adjustment* pada nilai *maximum number of generation*, yaitu dimulai dari 10-100 dengan kelipatan nilai 10, untuk menentukan jumlah generasi yang menghasilkan akurasi paling tinggi. Ketika *maximum of number generation* di-*adjustment*, nilai *pop size*, *p mutation*, dan *p crossover* berada pada nilai default, yaitu 5 untuk *pop size*, -1.0 untuk *p mutation*, dan 0.5 untuk *p crossover*. Setelah didapatkan jumlah maksimal generasi yang menghasilkan akurasi paling tinggi, kemudian dilanjutkan dengan melakukan *adjustment* pada nilai *pop size* yang dimulai dari 50-1000 dengan kelipatan nilai 50. Nilai pop size yang menghasilkan akurasi paling tinggilah yang akan digunakan pada langkah percobaan selanjutnya. Setelah itu, dilakukan *adjustment* pada *pc* dengan range 0.1-1.0 dengan kelipatan nilai 0.1. Dan yang terakhir dilakukan *adjustment* pada *pm* dengan range -1.0-1.0 dengan kelipatan nilai 0.1.

Berikut *adjustment* pada nilai *maximum number of generation:*

Tabel 4.1 *Adjustment* pada nilai *max number of generation*

Max num of generation	*Pop Size*	*P Crossover*	*P Mutation*	Jumlah Atribut	Akurasi	AUC
10	5	0.5	-1.0	4, nama parpol, kota administrasi, nomor urut caleg, suara sah caleg	96,78% +/- 0,77%	0,990 +/-0,010
20	5	0.5	-1.0	4, nama parpol, kecamatan, nomor urut caleg, suara sah caleg	96,78% +/- 0,77%	0,990 +/-0,010
30	5	0.5	-1.0	3, kota administrasi, nomor urut caleg, suara sah caleg	97,53% +/- 1,08%	0,984+/- 0,029
40	5	0.5	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,62% +/- 1,03%	0,993 +/- 0,005
50	5	0.5	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,62% +/- 1,03%	0,993 +/- 0,005
60	5	0.5	-1.0	3, jenis kelamin, kota administrasi, suara sah caleg	97,62% +/- 0,84%	0,993 +/- 0,005
70	5	0.5	-1.0	4, jenis kelamin, kota administrasi, daerah pemilihan, suara sah caleg	97,71% +/- 1,24%	0,993 +/- 0,003
80	5	0.5	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,75% +/- 1,00%	0,993 +/- 0,006
90	5	0.5	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,75% +/- 1,00%	0,993 +/- 0,006
100	5	0.5	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,75% +/- 1.00%	0,993 +/- 0,006

Akurasi tertinggi dicapai pada saat jumlah generasi 80 hingga 100. Nilai akurasi dan AUC dan jumlah atribut yang diperoleh adalah sama, yaitu berturut-turut 97.75% dan 0.993, 3 atribut (jenis kelamin, daerah pemilihan, suara sah caleg). Semakin banyak jumlah generasi, maka waktu komputasi yang dibutuhkan juga akan semakin bertambah. Sehingga diambil nilai terendah dari nilai-nilai jumlah generasi yang menghasilkan akurasi paling tinggi. yaitu 80. Nilai *maximum number of generation* yang digunakan untuk percobaan selanjutnya adalah 80. Selanjutnya dilakukan *adjustment* pada nilai *pop size* mulai dari 50-1000.

Tabel. 4.2 *Adjustment* pada nilai *pop size*

Max num of generation	*Pop Size*	*P Crossover*	*P Mutation*	Jumlah Atribut	Akurasi	AUC
80	50	0.5	-1.0	2, daerah pemilihan, suara sah caleg	97,79%	0,993
80	100	0.5	-1.0	2, daerah pemilihan, suara sah caleg	97,80%	0,995
80	150	0.5	-1.0	2, daerah pemilihan, suara sah caleg	97,80%	0,993
80	200	0.5	-1.0	2, daerah pemilihan, suara sah caleg	97,80%	0,994
80	250	0.5	-1.0	2, daerah pemilihan umum, suara sah caleg	97,80%	0,994
80	300	0.5	-1.0	2, daerah pemilihan umum, suara sah caleg	97,80%	0,994
80	350	0.5	-1.0	3, jenis kelamin, daerah pemilihan,	97,84%	0,994

				suara sah caleg		
80	400	0.5	-1.0	2, daerah pemilihan, suara sah caleg	97,80%	0,993
80	450	0.5	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,80%	0.994
80	500	0.5	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	550	0.5	-1.0	4, jenis kelamin, kota administrasi, daerah pemilihan, suara sah caleg	97,80%	0,994
80	600	0.5	-1.0	4, jenis kelamin, kota administrasi, daerah pemilihan, suara sah caleg	97,80%	0,993
80	650	0.5	-1.0	2, daerah pemilihan, suara sah caleg	97,80%	0,994%
80	700	0.5	-1.0	4, jenis kelamin, kota administrasi, daerah pemilihan, suara sah caleg	97,80%	0,993
80	750	0,5	-1,0	1, suara sah caleg	97,84%	0,994
80	800	0.5	-1.0	2, daerah pemilihan, suara sah caleg	97,80%	0,993
80	850	0.5	-1.0	2, daerah pemilihan, suara sah caleg	97,84%	0,993
80	900	0.5	-1.0	2, daerah pemilihan, suara sah caleg	97,80%	0,993
80	950	0.5	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	1000	0.5	-1.0	1, suara sah caleg	97,80%	0,993

Berdasarkan proses *adjustment* pada nilai *pop size* di atas, diperoleh akurasi tertinggi yang tercapai adalah 97,84% dengan nilai AUC 0,994. Akurasi ini dicapai oleh nilai *pop size* 350, 500, 750, dan 950. Semakin tinggi nilai *pop size* waktu komputasi yang dibutuhkan juga semakin lama. Karena ke-empat nilai *pop size* memiliki akurasi yang sama, maka dipilih nilai *pop size* terkecil yaitu 350 untuk eksperimen selanjutnya. Kemudian berikutnya akan dilakukan *adjustment* pada nilai *probability crossover* mulai dari 0.1 hingga 1.0.

Tabel 4.3 *Adjustment* pada nilai *probability crossover*

Max num of generation	*Pop Size*	*P Crossover*	*P Mutation*	Jumlah Atribut	Akurasi	AUC
80	350	0.1	-1.0	4, jenis kelamin, kota administrasi, daerah pemilihan, suara sah caleg	97,80%	0,993
80	350	0.2	-1.0	2, daerah pemiliihan, suara sah caleg	97,80%	0,993
80	350	0.3	-1.0	2, daerah pemiliihan, suara sah caleg	97,80%	0,994
80	350	0.4	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,80%	0,994
80	350	0.5	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.6	-1.0	4, jenis kelamin, kota administrasi, daerah pemilihan, suara sah caleg	97,80%	0,993
80	350	0.7	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,80%	0,993
80	350	0.8	-1.0	2, daerah pemilihan, suara sah caleg	97,80%	0,993
80	350	0.9	-1.0	1, suara sah caleg	97,80%	0,993
80	350	1.0	-1.0	2, daerah pemilihan, suara sah caleg	97,80%	0,993

Dari hasil *adjustment* nilai *pc* di atas, diperoleh nilai pc 0.5 yang dapat menghasilkan akurasi paling tinggi yaitu sebesar 97.84 % dengan nilai AUC 0,994. Kemudian langkah selanjutnya adalah melakukan *adjustment* pada nilai *probability mutation*. Berikut adalah hasil *adjustment* pada nilai *pm* dari -1.0 hingga 1.0:

Tabel 4.4 *Adjustment* pada probability mutation

Max num of generation	*Pop Size*	*P Crossover*	*P Mutation*	Jumlah Atribut	Akurasi	AUC
80	350	0.5	-1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	-0.9	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	-0.8	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	-0.7	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	-0.6	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	-0.5	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	-0.4	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	-0.3	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	-0.2	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	-0.1	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	0.0	2, daerah pemilihan, suara sah caleg	97,80%	0,993
80	350	0.5	0.1	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,80%	0,993
80	350	0.5	0.2	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,80%	0,994
80	350	0.5	0.3	2, daerah pemilihan, suara sah caleg	97,80%	0,993
80	350	0.5	0.4	2, daerah pemilihan, suara sah caleg	97,80%	0,993
80	350	0.5	0.5	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	0.6	2, daerah pemilihan, suara sah caleg	97,80%	0,994
80	350	0.5	0.7	2, daerah pemilihan, suara sah caleg	97,80%	0,994
80	350	0.5	0.8	2, daerah pemilihan, suara sah caleg	97,80%	0,993
80	350	0.5	0.9	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,84%	0,994
80	350	0.5	1.0	3, jenis kelamin, daerah pemilihan, suara sah caleg	97,79%	0,994

Dari percobaan di atas, terlihat bahwa nilai *pm* -1 hingga -0.1, 0.5, dan 0.9 menghasilkan akurasi paling tinggi, yaitu 97.84%, nilai AUC 0,994, dengan jumlah atribut 3 yaitu jenis kelamin, daerah pemilihan, suara sah caleg).

Dari proses fitur seleksi yang dilakukan algoritma genetika, maka atribut yang digunakan oleh naïve bayes untuk memprediksikan hasil pemilu legislatif DKI Jakarta adalah jenis kelamin, daerah pemilihan dan suara sah calon legislatif.

4.2 Evaluasi dan Validasi Hasil

4.2.1 Hasil Pengujian Metode Naïve Bayes

a. Confusion Matrix

Berdasarkan data training sebanyak 2267 record dengan 9 atribut diantaranya adalah nama parpol, nama caleg, no urut caleg, jenis kelamin, kota administrasi, daerah pemilihan, suara sah caleg, suara sah partai, jumlah perolehan kursi, yang dimodelkan dengan algoritma naïve bayes diperoleh hasil sebagai berikut:

accuracy: 92.28% +/- 1.91% (mikro: 92.28%)

	true TIDAK	true YA	class precision
pred. TIDAK	2006	8	99.60%
pred. YA	167	86	33.99%
class recall	92.31%	91.49%	

Gambar 4.6 Nilai akurasi model naïve bayes

Jumlah *true positive* (tp) adalah 2006 *record* diklasifikasikan sebagai TIDAK terpilih dan *False Negative* (fn) sebanyak 8 record diklasifikasikan sebagai TIDAK terpilih tetapi YA terpilih. Jumlah *true negative* (tn) adalah 86

record diklasifikasikan YA terpilih dan *false positive* (fp) sebanyak 167 record diklasifiaksikan YA terpilih tetapi TIDAK terpilih.

Dari *confusion matrix* di atas, terlihat bahwa akurasi dengan menggunakan algoritma naïve bayes adalah sebesar 92,28%. Berikut adalah perhitungan akurasi, sensitivity, specifity, *ppv*, dan *npv*.

$$acc = \frac{tp + tn}{tp + tn + fp + fn} = \frac{2006 + 8}{2006 + 8 + 167 + 8} = 0{,}9228$$

$$Sensitivity = \frac{tp}{tp + fn} = \frac{2006}{2006 + 8} = 0{,}9960$$

$$Specificity = \frac{tn}{tn + fp} = \frac{86}{86 + 167} = 0{,}3399$$

$$PPV = \frac{tp}{tp + fp} = \frac{2006}{2006 + 167} = 0{,}9231$$

$$NPV = \frac{tn}{tn + fn} = \frac{86}{86 + 8} = 0{,}9149$$

b. Evaluasi ROC curve

Gambar 4.7 Nilai AUC model naïve bayes

Berdasarkan grafik ROC diatas, terlihat bahwa nilai AUC (Area Under Curve) sebesar 0,981 dengan tingkat akurasi *Excellent Classification*.

4.2.2 Hasil Pengujian Naïve bayes dengan Algoritma Genetika sebagai Fitur Seleksi

a. Confusion Matrix

Berdasarkan data training sebanyak 2267 record dengan 9 atribut diantaranya adalah nama parpol, nama caleg, no urut caleg, jenis kelamin, kota administrasi, daerah pemilihan, suara sah caleg, suara sah partai, jumlah perolehan kursi, yang dimodelkan dengan algoritma naïve bayes dengan fitur seleksi menggunakan algoritma genetika diperoleh hasil sebagai berikut:

accuracy: 97.84% +/- 0.50% (mikro: 97.84%)

	true TIDAK	true YA	class precision
pred. TIDAK	2136	12	99.44%
pred. YA	37	82	68.91%
class recall	98.30%	87.23%	

Gambar 4.8 Nilai akurasi model naïve bayes dan algoritma genetika

Jumlah *true positive* (tp) adalah 2136 *record* diklasifikasikan sebagai TIDAK terpilih dan *False Negative* (fn) sebanyak 12 record diklasifikasikan sebagai TIDAK terpilih tetapi YA terpilih. Jumlah *true negative* (tn) adalah 82 record diklasifikasikan YA terpilih dan *false positive* (fp) sebanyak 37 record diklasifiaksikan YA terpilih tetapi TIDAK terpilih.

Dari *confusion matrix* di atas, terlihat bahwa akurasi dengan menggunakan algoritma naïve bayes adalah sebesar 97,84%. Berikut adalah perhitungan akurasi, sensitivity, specifity, *ppv*, dan *npv*.

$$acc = \frac{tp + tn}{tp + tn + fp + fn} = \frac{2136 + 82}{2136 + 82 + 37 + 12} = 0{,}9784$$

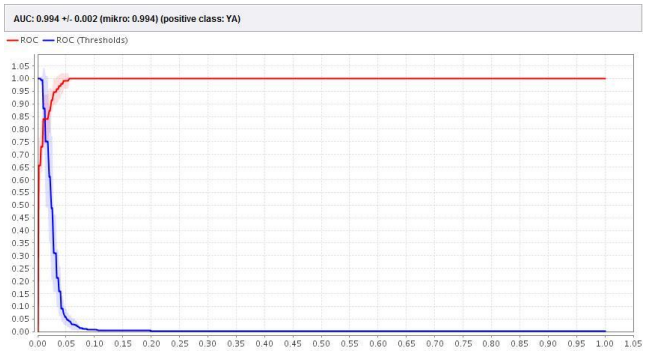
$$Sensitivity = \frac{tp}{tp + fn} = \frac{2136}{2136 + 12} = 0{,}9944$$

$$Specificity = \frac{tn}{tn + fp} = \frac{82}{82 + 37} = 0{,}6891$$

$$PPV = \frac{tp}{tp + fp} = \frac{2136}{2136 + 37} = 0{,}9830$$

$$NPV = \frac{tn}{tn + fn} = \frac{82}{82 + 12} = 0{,}8723$$

b. Evaluasi ROC curve



Gambar 4.9 Nilai AUC model naïve bayes dan algoritma genetika

Berdasarkan grafik ROC diatas, terlihat bahwa nilai AUC (Area Under Curve) sebesar 0,994 dengan tingkat akurasi *Excellent Classification*.

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Algoritma naïve bayes dengan algoritma genetika sebagai fitur seleksi dan algoritma naïve bayes tanpa fitur seleksi, dapat diterapkan untuk prediksi pemilu legislatif DKI Jakarta. Keduanya menghasilkan akurasi yang *excellent*. Prediksi hasil pemilu legislatif DKI Jakarta dengan algoritma naïve bayes tanpa fitur seleksi memiliki akurasi 92,28% dengan nilai AUC 0,981, sedangkan prediksi hasil pemilu legislatif DKI Jakarta menggunakan algoritma naïve bayes dengan algoritma genetika sebagai fitur seleksi memiliki akurasi 97.84% dengan

nilai AUC 0,994. Jadi prediksi pemilu legislatif DKI Jakarta menggunakan algoritma naïve bayes dengan algoritma genetika sebagai fitur seleksi lebih unggul dibandingkan menggunakan algoritma naïve bayes tanpa fitur seleksi.

5.2 Saran

Agar penelitian ini terus berkembang, berikut saran-saran yang diusulkan :

1. Penelitian ini dapat dikembangkan dengan menggunakan metode-metode klasifikasi data mining lainnya untuk melakukan perbandingan.
2. Peneltian ini dapat dikembangkan dengan menggunakan algoritma fitur seleksi atau dengan algoritma optimasi lainnya.
3. Penelitian ini diharapkan dapat digunakan oleh masyarakat dan para politikus untuk mengetahui kebijakan apa yang harus diterapkan terkait dengan pemilu legislatif DKI Jakarta.

REFERENSI

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

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

[3] *Choi, J. H., & Han, S. T. (1999). Prediction of Elections Result using Descrimination of Non-Respondents:The Case of the 1997 Korea Presidential Election.*

[4] *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 .*

[5] *Moscato, P., Mathieson, L., Mendes, A., & Berreta, R. (2005). The Electronic Primaries:Prediction The U.S. Presidential Using Feature Selection with safe data. ACSC '05 Proceeding of the twenty-eighth Australian conference on Computer Science Volume 38 , 371-379.*

[6] *Zhang, H., dan Su, J. (2007). Naive bayesian classi_ers for ranking. Retrieved December 2007, from www.cs.unb.ca/profs/hzhang/publications/NBRanking.*

[7] *Santosa, B. 2007. Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis. Graha Ilmu:Yogyakarta.*

[8] *Davies, and Paul Beynon, 2004, "Database Systems Third Edition", Palgrave Macmillan, New York.*

[9] *Han, J. and Kamber, M, 2006, "Data Mining Concepts and Techniques Second Edition". Morgan Kauffman, San Francisco.*

[10] *Witten, I. H and Frank, E. 2005. Data Mining : Practical Machine Learning Tools and Techniques Second Edition. Morgan Kauffman : San Francisco.*

[11] *Prasetyo, Eko. 2012. Data Mining :Konsep dan Aplikasi menggunakan Matlab. Penerbit Andi: Yogyakarta.*

[12] *Jananto, Arief. 2010. Memprediksi Kinerja Mahasiswa Menggunakan Teknik Data Mining (Studi kasus data akademik mahasiswa UNISBANK. Tesis Tidak Terpublikasi. Yogyakarta: Universitas Gajah Mada.*

[13] *Badrul, Mohammad. 2012. Prediksi Hasil Pemilu Legislatif Dki Jakarta Dengan Metode Neural Network Berbasis Particle Swarm Optimization. Tesis Tidak Terpublikasi. Jakarta: STMIK Nusa Mandiri.*

[14] *Manurung, Adler Haymans, Cahyanti Ira K.2006. Pengaruh Peristiwa Politik (Pengumuman Hasil Pemilu Legislatif, Pengumuman Hasil Pemilihan Presiden, Pengumuman Susunan Kabinet, Reshuffle Kabinet) Terhadap Sektor- Sektor Industri Di Bursa Efek Jakarta. Jakarta:*

[15] *Sartono, Bagus. 2010. Pengenalan Algoritma genetic untuk Pemilihan Peubah Penjelas dalam Model Regresi Menggunakan SAS/IML. Forum Statistika dan Komputasi, Oktober 2010 p:10-15 Vol 15 No 2.*

[16] *Admin. 2009. IPTEK VOICE : Metode Quick Count. Dikutip dari http://www.ristek.go.id , diakses pada tanggal 26 November 2013.*

[17] *Permata Sari, Eka, Rini Sovia, M.Kom , Mardison, M.Kom. 2009. Algoritma Genetika Untuk Optimasi Pengaturan Jadwal Kuliah. Skripsi Tidak Terpublikasi. Universitas Putra Indonesia YPTK : Padang.*

[18] *Kusumadewi,Sri.2003.Artificial Intelegence (Teknik dan Aplikasinya).Yogyakarta:Graha Ilmu.*

[19] *Fauzi Rahman, Riza. 2011. Optimalisasi Antrian Pembelian Karcis Di Stasiun Bandung Dengan Menggunakan Algoritma Genetika.Skripsi. Jakarta:UPI*

[20] *Maimon, O., & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook. London: Springer.*

[21] *Salappa, A., Doumpos, M., & Zopounidis, C. (2007). Feature Selection Algorithms in Classification Problems: An Experimental Evaluation. Systems Analysis, Optimization and Data Mining in Biomedicine , 199-212.*

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

[23] *Han, J., & Kamber, M. (2007). Data Mining Concepts and Technique. Morgan Kaufmann publisher.*

[24] *Bramer, Max. (2007). Principles of Data Mining. London: Springer. ISBN-10: 1-84628-765-0, ISBN-13: 978-1-84628-765-7.*

[25] *Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. 2000. CRISP-DM 1.0 : Step-by-Step Data Mining Guide. Tersedia di*

[26] *http://www.community.udayton.edu/provost/it/training/documents/SPSS_CRISPWPlr.pdf. [diunduh : 10 Desember 2010].*

[27] *Sardini, N. H. (2011). Restorasi penyelenggaraan pemilu di Indonesia. Yogyakarta: Fajar Media Press.*

[28] *Undang-Undang RI No.10. (2008).*

[29] *Suyanto, MT, Msc. 2007. Artificial Intelligent, Searching, Reasoning, Planning dan Learning. Bandung : Informatika Bandung.*

[30] *Anita,Desiani, Arhami Muhammad. 2006. Konsep Kecerdasan Buatan. Yogyakarta: Cv. Andi Offset.*

[31] *Kusumadewi, Sri. 2003. Artificial Intelligent. Yogyakarta : Graha Ilmu.*

[32] *Binus.http://library.binus.ac.id/eColls/eThesisdoc.[diunduh: 25 Februari 2014]*

[33] *Han, J., & Kamber, M. (2006). Data Mining Concept and Tehniques. San Fransisco: Morgan Kauffman. ISBN 13: 978-1-55860-901-3*

[34] *Powers, D.M.W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness& Correlation. Journal of Machine Learning Technologies,ISSN: 2229-3981 & ISSN: 2229-399X, Volume 2, Issue 1, 2011, pp-37-63*

[35] *Gorunescu, F. (2011). Data Mining Concept Model and Techniques. Berlin: Springer. ISBN 978-3-642-19720-8*