Perbandingan Performansi Algoritma *Decision Tree* CART dan CHAID

Mohamad Abdul Kadir / 13507134
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
adi.stei@itb.ac.id

Kasus prediksi resiko kredit dan prediksi hipertensi esensial merupakan kasus yang dapat dibantu dengan teknik-teknik data mining. Sudah banyak teknik-teknik data mining yang diterapkan untuk membantu analisis prediksi kedua kasus tersebut. Dalam paper ini, yang akan disorot adalah penggunaan teknik data mining Decision Tree CART dan CHAID. Model tree untuk kedua algoritma tersebut agak berbeda untuk data set yang sama. Tingkat akurasi untuk prediksi resiko kredit masing-masing sebesar CHART 87,27% dan CHAID 87,15%. Sedangkan untuk kasus kedua akurasi dihitung berdasarkan sensitivity, specificity, dan PR (predictive rate).

Kata kunci: Algoritma, CHAID, CHART, *data mining*, *decision tree*, hipertensi, kredit, pohon, prediksi

I. PENDAHULUAN

Kemajuan teknologi informasi telah menyebabkan banyak orang dapat memperoleh data dengan mudah bahkan cenderung berlebihan. Data tersebut semakin lama semakin banyak dan terakumulasi, akibatnyapemanfaatan data yang terakumulasi tersebut menjadi tidak optimal.

Sebagai contoh perusahaan *retail* akan memberikan brosur penawaran barang-barang yang dijual ke pelanggan sesuai basis data pelanggan yang mereka punya. Jika perusahaan *retail* tersebut mempunyai 1.000.000 data pelanggan dan masing-masing pelanggan tersebut dikirimkan sebuah brosur penawaran dimana biaya pengiriman brosur tersebut adalah Rp 2.000,00 , maka biaya yang akan dikeluarkan oleh perusahaan tersebut adalah Rp 2.000.000.000,00 per bulan. Dari penggunaan dana tersebut mungkin hanya sepertiganya atau bahkan 8% saja yang secara efektif membeli penawaran tersebut.

Sehingga diperlukan analisis nasabah yang potensial membeli produk tertentu dan membedakan pengiriman brosur sesuai dengan potensi pembelian dari pelanggan

Data mining adalah salah satu solusi untuk permasalahan di atas. *Data mining* merupakan serangkaian proses untuk menggali suatu informasi terpendam dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. *Data mining* akan membentuk klasifikasi dalam kelompok yang memiliki karakteristiknya masing-masing. Proses klasifikasi kelompok ini biasa disebut teknik *data mining* atau algoritma *data mining* [FAR07].

Terdapat teknologi *data mining* yang telah telah dikembangkan diantaranya *clustering, classification, association rule, neural network, decision tree,* dan lain-lain. Tapi bagaimana memilih teknik *data mining* yang tepat sehingga dihasilkan klasifikasi dan prediksi yang akurat?

Pada tulisan ini, penulis akan melakukan kajian perbandingan performansi algoritma *decision tree* CART dan CHAID. Keduanya akan dibandingkan untuk 2 studi kasus, yakni kasus prediksi status resiko kredit di bank tertentu melalui *credit scoring* dan kasus prediksi hipertensi esensial.

II. *DECISION TREE*

Decision tree merupakan salah satu teknik yang dapat digunakan untuk melakukan klasifikasi terhadap sekumpulan objek atau *record*. Teknik ini terdiri dari kumpulan *decision node*, dihubungkan oleh cabang, bergerak ke bawah dari *root node* sampai berakhir di *leaf node*. Pengembangan *decision tree* dimulai dari *root node*, berdasarkan konvensi ditempatkan di bagian atas diagram *decision tree*, semua atribut dievaluasi pada *decision node*, dengan tiap *outcome* yang mungkin menghasilkan cabang. Tiap cabang dapat masuk baik ke *decision node* yang lain ataupun ke *leaf node* [YUS07].

Decision tree adalah metode diskrimasi nonlinear yang menggunakan sekumpulan variabel independen untuk membagi sampel ke dalam kelompok–kelompok yang lebih kecil secara bertahap. Prosedur tersebut dilakukan secara iterative di setiap cabang pohon, yakni memilih variabel independen yang memiliki asosiasi terkuat dengan variabel dependen menurut kriteria tertentu.

Persyaratan yang harus dipenuhi dalam penerapan algoritma *decision tree* [YUS07]:

1. Algoritma *decision tree* merepresentasikan *supervised learning* sehingga membutuhkan target *preclassified.*
2. *Training data set* harus kaya dan bervariasi.
3. Kelas atribut target harus diskrit.

III. CART (*CLASSIFICATION AND REGRESSION TREES*)

CART merupakan metode partisi rekursif yang digunakan baik untuk regresi maupun klasifikasi. CART dibangun dengan melakukan pemecahan subset-subset dari dataset menggunakan variabel prediktor untuk membuat dua *child node* secara berulang, dimulai dari keseluruhan dataset. Tujuannya adalah menghasilkan subset data yang sehomogen mungkin untuk mengklasifikasikan variabel target.

Pada permulaan proses, *training set* yang terdiri *record* yang sudah diklasifikasi harus tersedia. *Training set* digunakan untuk membangun *tree* yang memungkinkan penempatan suatu kelas ke dalam variabel target dari *record* baru yang didasarkan pada nilai-nilai variable yang lain atau variable independen [YUS07].

CART membangun *binary tree* dengan memecah *record* pada tiap *node* berdasarkan fungsi variable input tunggal. Tugas pertama yang dijalankan adalah menentukan variable independen yang menjadi *splitter* terbaik. *Splitter* terbaik adalah *splitter* yang menurunkan keanekaragaman *node. Node* yang tidak dipecah lagi disebut *leaf node* [YUS07].

Pemecahan *record* pada tiap *node* menyebabkan jumlah *record* yang semakin kecil dari *root node* ke *child node* sampai ke *leaf node.* Semakin sedikit jumlah *record*, semakin kurang representative *node* tersebut. Akibatnya adalah model *tree* hanya dapat memprediksi secara akurat untuk *record* yang berada pada *training set*, tetapi tidak dapat memprediksi *record* baru yang berasal dari luar *training set* secara akurat atau*overtraining.* Untuk mengurangi *overtraining,* pemangkasan pohon atau *pruning* dapat dilakukan. *Pruning* menghasilkan beberapa kandidat *subtree* [YUS07].

Beberapa kandidat *subtree* dipilih berdasarkan kemampuannya dalam memprediksi *record* baru. Pemilihan tersebut membutuhkan set data bari yaitu set *test set* yang berisi *record* baru yang berbeda dengan *record* yang ada pada *training set*. Tiap kandidat *subtree* digunakan untuk memprediksi *record* yang ada dalam *test set. Subtree* yang memberikan *error* terkecil terpilih sebagai model *tree* [YUS07].

Langkah terakhir adalah mengevaluasi *subtree* terpilih dengan menerapkannya pada *set data* baru yaitu *validation set.* Nilai *error* yang diperoleh dari *validation set* digunakan untuk memprediksi *expected performance* model prediksi [YUS07].

IV. CHAID (CHI-SQUARED AUTOMATIC INTERACTION DETECTION)

Pembangunan *tree* dengan CHAID berbeda dengan CART. CART membangun *tree* dengan *overfitting* data, kemudian melakukan *pruning,* CHAID akan menghentikan pembangunan *tree* sebelum *overfitting* terjadi [YUS07].

Metode CHAID adalah berdasarkan tes *chi-square* terhadap asosiasi. Pohon CHAID adalah *decision tree* yang dibangun dengan memecah/*splitting* subset-subset secara berulang ke dalam dua atau lebih *child node* yang dimulai dari keseluruhan dataset. Metode CHAID pada dasarnya berhubungan dengan interaksi-interaksi antara variabel independen yang tersedia secara langsung dari pengujian pohon.

Tiap variabel prediktor dipertimbangkan sebagai *splitter*. Tahap pertama dalam investigasi ini adalah menggabungkan kategori-kategori yang berkorespondensi dengan nilai variable target yang sama. Seluruh variable predictor yang tidak menghasilkan perbedaan yang signifikan dalam nilai variable target digabung [YUS07].

Dalam tahap kedua, tiap grup dari tiga atau lebih predictor dipecah kembali dengan seluruh pembagian biner yang mungkin. Jika pemecahan ini menghasilkan perbedaan yang signifikan maka pemecahan tersebut dipertahankan [YUS07].

Setelah tiap variabel prediktor dikelompokkan untuk menghasilkan keanekaragaman kelas yang maksimum dalam variable target, tes x^2 diterapkan pada kelompok tersebut. Prediktor yang menghasilkan kelompok-kelompok paling berbeda dipilih sebagai *splitter* pada *node* tersebut [YUS07].

V. KASUS PREDIKSI STATUS RESIKO KREDIT BANK MELALUI *CREDIT SCORING*

Bank mempunyai peranan yang esensial dalam penyaluran kredit kepada pihak-pihak yang membutuhkan. Fungsi pokok kredit yaitu memenuhi pelayanan terhadap kebutuhan masyarakat dalam rangka memperlancar perdagangan, produksi, dan jasa-jasa bahkan konsumsi yang kesemuanya itu ditujukan untuk meningkatkan kesejahteraan manusia. Salah satu unsur dalam kredit adalah janji dan kesanggupan membayar dari debitur [FIR04].

Pembayaran hutang dari debitur kepada kreditur tidak selamanya mulus. Seringkali terjadi kredit macet.

Perkembangan teknologi informasi telah mempengaruhi cara penilaian resiko yang semula dengan cara *human judgment* bergeser ke arah yang formal dan objektif, yaitu melalui *credit scoring.* Tujuan dari *credit scoring* adalah membantu pihak kreditur mengkuantifikasi resiko financial sehingga keputusan danpat diambil cepat dan lebih akurat [CHY04].

Banyak teknik yang dapat membantu dalam pembangunan model *credit scoring,* salah satunya dengan teknik *data mining. Decision tree* merupakan salah satu teknik *data mining* yang popular karena mudah diinterpretasikan dan divisualisasikan [CHY04].

Beberapa algoritma *decision tree* yang dapat digunakan untuk membangun model *tree* diantaranya adalah CART dan CHAID. Keduanya menghasilkan model *tree* dan keakuratan berbeda untuk set data yang sama

V. A. *Credit Scoring*

Credit scoring adalah *tool* yang melibatkan

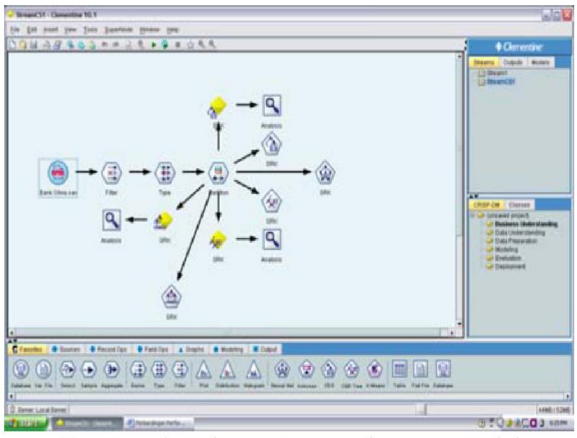
penggunaan model statistic untuk mengevaluasi seluruh informasi yang tersedia secara objektif dalam pengambilan keputusan kredit [NOE97]. Manfaatnya adalah:

- Peningkatan kecepatan dan konsistensi proses aplikasi pinjaman dan memungkinkan otomatisasi proses peminjaman;
- Adanya kemampuan belajar sepanjang waktu karena model ini didasarkan pada perhitungan statistic data masa lalu.

Model ini dibangun dengan menggunakan sampel kredit masa lalu dalam jumlah yang besar. Sampel tersebut dibagi ke dalam dua kelas yaitu kredit yang baik dan kredit yang bermasalah. Berdasarkan pola masa lalu, kombinasi karakteristik peminjam yang membedakan peminjam yang baik dan yang buruk menghasilkan nilai sebagai estimasi resiko tiap peminjam baru [YUS07].

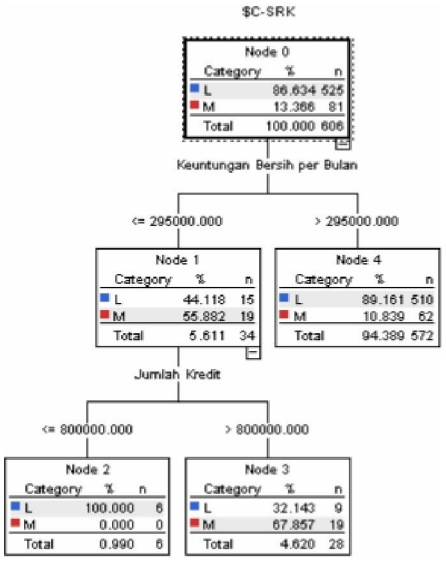
V. B. Pengembangan model *tree*

Model *tree* yang dikembangkan ialah model *tree* untuk memprediksi status resiko kredit. Set data terdiri dari 11 variabel predictor yaitu jenis kelamin, umur, jumlah kredit, lama pinjaman, angsuran per bulan, total angsuran per bulan, jenis pengajuan, sector ekonomi, omzet per bulan, keuntungan bersih per bulan, jaminan, dan nilai jaminan; dan satu variable target yaitu status resiko kredit. Pengembangan model *tree* ini dibantu dengan Clementine.



Gambar 1. Tahapan pengembangan Model *Decision Tree* [YUS07]

Model *tree* yang dihasilkan sangat bergantung pada komposisi *record* pada tiap set data *training*, *test*, dan *validation.* Untuk menghindari bias, pengembangan model dilakukan sebanyak 20 kali dengan komposisi *record* pada setiap set data yang berbeda dari satu pengembangan ke pengembangan lainnya.



Gambar 2. Model *Tree* Status Resiko Kredit [YUS07]

Tabel 1. Performansi Algoritma *Decision Tree* [YUS07]

Model	CART	CHAID
1	86,86	84,62
2	88,04	89,7
3	86,27	83,33
4	88,52	88,85
5	90,15	88,32
6	87,66	86,69
7	87,29	86,26
8	83,85	84,16
9	86,97	87,3
10	87,66	86,69
11	87,14	88
12	87,58	87,58
13	84,82	86,47
14	86,14	92,13
15	90,24	90,24
16	87,38	85,05
17	86,26	87,54
18	88,6	88,93
19	85,46	86,17
20	87,63	84,95

V. C. Analisis perbandingan

Algoritma CART menghasilkan rata-rata tingkat keakuratan sebesar 87,28% dan standar deviasi 1,51 sedangkan CHAID rata-rata sebesar 87,15% dan standar deviasi 2,19.

Analisis variansi dilakukan untuk menguji apakah ada perbedaan variansi secara signifikan di antara ketiga algoritma tersebut. Hasil analisis menunjukkan bahwa tidak ada perbedaan variansi secara signifikan.

Tabel 2. Analisis Variansi [YUS07]

Tabel 2. Analisis variansi

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Antar Algoritma	3.622	2	1.811	.571	.568
Dalam Algoritma	180.710	57	3.170		
Total	184.332	59			

Selain uji variansi, dilakukan juga perbandingan rata-rata yang memberikan hasil bahwa tidak ada perbedaan rata-rata secara signifikan di antara kedua algoritma tersebut.

Tabel 3. Perbandingan rata-rata performansi

Algoritma (I)	Algoritma (J)	Mean Diff (I-J)	Sig.
CART	CHAID	.12550	.824
CHAID	CART	-.12550	.824

VI. KASUS PREDIKSI HIPERTENSI ESENSIAL

Hipertensi adalah faktor resiko yang signifikan terhadap penyakit jantung, stroke, gagal jantung, penyakit ginjal, dan beberapa masalah kardiovaskular. Terdapat banyak factor yang berhubungan dengan penyakit hipertensi ini, diantaranya adalah usia, jenis kelamin, riwayat hipertensi keluarga, kebiasaan merokok, lipoprotein, asam urat, jumlah kolesterol dan *body mass index* (BMI).

Prediksi hasil dari penyakit adalah satu dari tugas menarik dan menantang untuk dikembangkannya aplikasi *data mining.* beberapa pekerjaan tentang perbandingan teknik-teknik klasifikasi dalam berbagai area telah dipublikasikan.

Pada kasus ini teknik klasifikasi yang digunakan adalah CART, MARS, CHAID, ID3, dan *artificial neural network.* Namun yang akan menjadi sorotan pada *paper* ini adalah teknik CART dan CHAID yang digunakan untuk kasus ini.

Tools yang digunakan dalam menganalisis teknik-teknik tersebut adalah **Answer Tree 2.1** [TUR05].

VI. A. Subjek dan dataset

Pada studi kasus ini, analisis retrospektif dilakukan pada 694 subjek (452 pasien dan 242 kontrol). Diagnosis hipertensi di buat saat rata-rata dari tiga atau lebih pengukuran diastolic *blood pressure* (BP) pada setidaknya tiga kali kunjungan adalah >= 90 mmHg atau saat rata-rata pembacaan sistolik BP pada tiga atau lebih kunjungan adalah >= 140 mmHg secara konsisten.

Variabel-variabel independen diantaranya usia, jenis kelamin, riwayat hipertensi keluarga, kebiasaan merokok, lipoprotein (a), trigliserid, asam urat, jumlah kolesterol, dan BMI. Sebelum membangun model, data *file* secara acak di-*split* menjadi dua dataset, 75% (n=520) *data training* dan 25% (n=174) untuk *test set*. Data tersebut dikoleksi dan diambil dari klinik kardiologi Fakultas Kedokteran Universitas Trakya di Turki.

VI. B. *Hierarchical Cluster Analysis* (HCA)

HCA adalah metode statistik untuk menemukan kluster-kluster homogen relatif dari kasus-kasus (dalam hal ini algoritma-algoritma data mining) berdasarkan karakteristik pengukuran. Analisis ini dimulai dari setiap kasus dalam kluster terpisah dan mengkombinasikan kluster-kluster secara sekuensial, mengurangi jumlah kluster dalam setiap tahapan hingga hanya satu kluster yang tersisa. Ketika terdapat N kasus, maka akan terjadi N-1 tahapan clustering atau fusi. Proses ini dapat direpresentasikan sebagai pohon atau dendrogram yang mana dalam setiap tahapan proses clustering diilustrasikan dengan *join* pohon. Pada *paper* ini tidak akan ditunjukkan pemrosesan clustering ini maupun hasil keseluruhan dari seluruh algoritma yang diklustering karena hanya akan difokuskan pada algoritma CHAID dan CART.

VI. C. Hasil

VI. C. 1. Karakteristik Pembanding

Karakteristik dari populasi subjek dapat dilihat pada Tabel 4. Pada penelitian ini akan ditunjukkan analisis statistic klasik untuk menguji perbedaan distribusi variabel antara hipertensi dan kelompok kontrol. Variabel numeric diuji untuk distribusi normal dengan uji Kolmogorov-Smirnov. *Age, total cholesterol,* dan BMI diuji dengan sampel independen *t-test*, tapi *lipoprotein, triglyseride,* dan *uric acid* diuji dengan Mann-Whitney *U-test* karena distribusi variabel-variabel tersebut nonnormal. Variabel nominal diuji dengan uji *chi-square* untuk hipertensi dan kelompok kontrol [TUR05].

Tabel 4. Karakteristik subjek studi [TUR05]

Characteristics	Hypertension ($n=452$)	Control ($n=242$)	p
Age (years)	48.2±8.6	46.5±8.2	0.015
Sex (female/male)	2.54	1.71	0.024
Family history of hypertension (%)	78.0	34.2	<0.001
Smoking habits (%)	76.1	67.5	0.019
Lipoprotein (a) (mg/dl)	26.0 (15.0–43.3)	21.4 (14.9–32.6)	0.006
Triglyceride (mg/dl)	146.0 (110.0–201.5)	111.0 (87.8–162.8)	<0.001
Uric acid (μmol/L)	4.1 (3.2–5.2)	3.4 (2.9–3.9)	<0.001
Total cholesterol (mmol/L)	210.0±45.3	200.1±40.0	0.004
BMI (kg/m^2)	28.8±3.6	27.3±3.6	<0.001

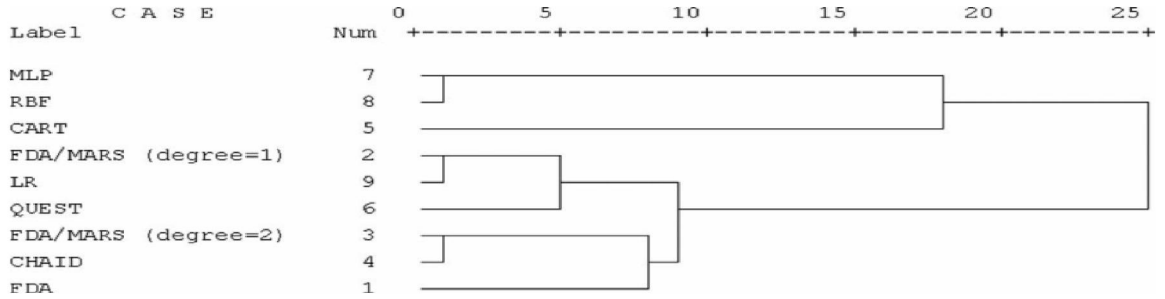
VI. C. 2. Perbandingan teknik-teknik klasifikasi

Perbandingan *sensitivity, specificity,* dan *predictive rate* (PR) untuk *training set* dan *test set* teknik klasifikasi ditunjukkan pada Tabel 5.

Sensitivity, specificity, dan PR untuk data training digunakan sebagai variabel input dalam HCA. HCA dilakukan untuk mengidentifikasi kelompok homogen dari teknik klasifikasi berdasarkan *sensitivity, specificity,* dan PR.

Tabel 5. Perbandingan performansi untuk *training set* dan *test set*

Model	Training set			Test set		
	Sensitivity (%)	Specificity (%)	PR (%)	Sensitivity (%)	Specificity (%)	PR (%)
CHAID	93.00	60.60	80.95	88.24	70.00	83.10
CART	87.30	70.20	80.95	82.35	70.00	78.87



Gambar 3. *Dendrogram* berikut ini menunjukkan relasi antar-teknik klasifikasi (di penelitian yang menjadi referensi, diuji beberapa teknik algoritma).

Representasi pohon CART dan CHAID cukup dekat/mirip dengan *medical reasoning* dan dapat membantu membentuk pemahaman terhadap prediksi. Metode-metode ini memiliki potensi untuk melengkapi model statistik yang sudah ada dan berkontribusi dalam membantu *decision support* yang terkomputerisasi dalam hal intepretasi dan presentasi resiko. Model-model ini menyediakan *framework* analisis komprehensif untuk menginformasikan rancangan optimal pedoman klinis dan kebijakan kesehatan untuk pencegahan dan manajemen hipertensi.

VII. PENUTUP

Algoritma CART dan CHAID memiliki performansi, yaitu tingkat akurasi dalam prediksi status resiko kredit yang tidak begitu jauh berbeda.

Model *tree* yang dibangun sangat bergantung pada komposisi *record* dalam *training set* dan *testing set*, untuk itu diperlukan pengujian perbandingan performansi di antara model-model *tree* yang dihasilkan.

Untuk kasus kedua, representasi pohon dalam CHAID dan CART cukup dekat/mirip dengan *medical reasoning* dan dapat membantu dalam pemahaman prediksi sehingga berkontribusi dalam interpretasi dan presentasi resiko untuk sistem *decision support* yang sudah terkomputerisasi.

REFERENSI

[CHY04] Cyhe, K.H., Chin, T.W., dan Peng, G.C., 2004. *Credit Scoring Using Data Mining Techniques*. Singapore Management Review 26 (2): 25-47.

[FAR07] Farisi, A..(2007). Perbandingan tigkat Akurasi Dua Model Data Mining yang dihasilkan oleh Decision Tree dan Naïve Bayes Studi Kasus: Suatu Perusahaan Manufaktur dan Penjualan Sepeda. Universitas Indonesia. Jakarta

[FIR04] Firdaus, Rachmat H., dan Maya Arianti, 2004. Manajemen Perkreditan Bank Umum. Bandung: Alfabeta.

[NOE97] Noe, J., 1997. *Credit Scoring*. America's Community Banker 6 (8): 29-33.

[TUR05] Ture, M., Kurt, I., Turhankurum, a., & Ozdamar, K. (2005). *Comparing classification techniques for predicting essential hypertension*. *Expert Systems with Applications*, *29*(3), 583-588. doi: 10.1016/j.eswa.2005.04.014.

[YUS07] Yusuf W., Yogi. (2007). PERBANDINGAN PERFORMANSI ALGORITMA DECISION TREE C5.0 , CART DAN CHAD: KASUS PREDIKSI STATUS RESIKO KREDIT DI BANK X ,. *Seminar*, *2007*(Snati), 0-3. UNPAR. Bandung