SISTEM PENDUKUNG KEPUTUSAN UNTUK MENENTUKAN PENJURUSAN SLTA DENGAN METODE ID3 DAN C4.5.

Eza Rahmanita1, Yeni Kustiyahningsih2
1. Universitas Trunojoyo, Madura
2. Univesitas Trunojoyo, Madura
Kontak Person:
Yeni Kustiyahningsih
JL. Raya Telang – PO.BOX 2 Kamal Bangkalan
Telp: 083849492078, E-mail: ykustiyahningsih@yahoo.com

ABSTRAK

Penjurusan sekolah dalam jenjang SLTA merupakan permasalahan yang komplek, hal ini dikarenakan jumlah siswa yang terus bertambah dan banyaknya syarat yang harus dipenuhi dalam menentukan jurusan, akibatnya proses penjurusan kurang tepat dan tidak sesuai minat siswa. Jika penjurusan sesuai dengan kemampuan dan minat/bakat siswa maka merea dapat belajar dengan nyaman dan dapat menghasilkan lulusan yang kompeten sesuai bidangnya. Tujuan penelitian ini adalah membantu guru dalam proses penyeleksian pemilihan jurusan sehingga proses yang dihasilkan dari seleksi ini lebih akurat dan objektif. Indikator atau kriteria yang akan digunakan dalam penyeleksian ini adalah nilai Matematika, Fisika, Biologi, Kimia untuk semester 1 dan semester 2, Nilai Psikotest (IQ), Saran Psikotest, Angket/Minat Siswa, Saran Bimbingan Konseling. Metode yang digunakan dalam penelitian ini adalah metode klasifikasi pohon keputusan ID3 dan C4.5. karena memiliki tingkat akurasi yang tinggi dalam menentukan keputusan. Hasil dari klasifikasi kedua algoritma akan di analisa untuk menentukan algoritma mana yang paling optimal kinerjanya. Kedua algoritma ini akan dibandingkan kinerjanya dengan mencari *recall*, *pressicion*, *accuracy* terbesar dan Nilai *error rate* terkecil yang dicapai. Hasil akhir dari penelitian ini, bahwa kinerja algoritma C4.5 lebih baik dari pada algoritma ID3 karena memiliki tingkat akurasi yang lebih tinggi dari pada algoritma ID3.

Kata Kunci : Algoritma ID3, C4.5, Penentuan Jurusan SLTA, Pohon keputusan, akurasi

1. PENDAHULUAN

Sekolah adalah tempat terjadinya proses belajar mengajar siswa dimana sekolah menyelenggarakan pendidikan yang berperan dalam pengembangan ilmu pengetahuan dan pengabdian kepada masyarakat. Sekolah lanjut tingkat atas (SLTA), adalah jenjang pendidikan menengah pada pendidikan formal di Indonesia setelah lulus Sekolah Menengah Pertama. Pada saat kelas XI, siswa SLTA dapat memilih jurusan yang ada. Idealnya, pemilihan jurusan itu berdasarkan minat, bakat, dan kemampuan siswa, sehingga dengan itu mereka diharapkan akan berhasil dalam menyelesaikan studinya di SLTA serta dapat melanjutkan pendidikan ke jenjang yang lebih tinggi. Penjurusan akan disesuaikan dengan minat dan kemampuan siswa, tujuannya agar pelajaran yang diberikan kepada siswa menjadi lebih terarah. Penjurusan yang tersedia di SLTA meliputi ilmu pengetahuan alam (IPA), ilmu pengetahuan sosial (IPS) dan bahasa. Terdapat 2 jurusan yaitu IPA dan IPS, penentuan penjurusan ini dipertimbangkan berdasarkan nilai akademik siswa, minat siswa dan bakat siswa yang dilihat dari hasil psikotest.

Proses penjurusan dilakukan pada saat siswa berada di kelas X dan akan naik ke kelas XI. Setelah wali kelas menerima seluruh nilai semester maka wali kelas akan memutuskan apakah siswa tersebut naik atau tidak. Jika siswa tersebut dinyatakan naik maka selanjutnya akan dilakukan proses penjurusan oleh tim yang terdiri dari wakil kepala sekolah bidang kurikulum, guru bimbingan konseling, wali kelas x dan guru mata pelajaran yang berkaitan dengan penjurusan. masalah yang sering terjadi dalam proses penjurusan adalah keterlambatan nilai siswa dari para wali kelas, akibatnya pada akhir proses penjurusan para tim penentu jurusan berburu waktu sehingga proses penjurusan kurang tepat, ditambah lagi dengan banyaknya jumlah siswa kelas X. Dari permasalahan tersebut maka penelitian ini menbangun aplikasi untuk menentukan penjurusan siswa SLTA metode yang di gunakan adalah metode klasifikasi pohon keputusan ID3 dan C4.5. Hasil dari klasifikasi kemudian akan dibandingkan untuk menentukan algoritma yang paling optimal kinerjanya. Tujuan penelitian ini adalah membuat aplikasi sistem pendukung keputusan untuk penjurusan siswa SLTA dan memberikan masukkan kepada pihak sekolah untuk memecahkan masalah yang muncul pada penjurusan siswa serta sebagai bahan pertimbangan guru BK dalam menentukan jurusan IPA dan IPS.

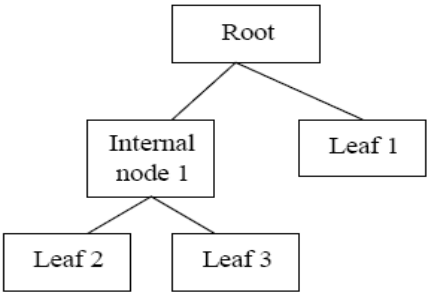
2. KAJIAN PUSTAKA

2.1. Pohon Keputusan (*Decision Tree*)

Salah satu metode *data mining* yang umum digunakan adalah pohon keputusan. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan *rule*. Pohon keputusan adalah salah satu metode klasifikasi yang paling popular karena mudah untuk diinterpretasi oleh manusia. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (Larose, 2005).

Gambar 2.1. Konsep pohon keputusan

Pohon keputusan merupakan himpunan aturan IF...THEN. Setiap *path* dalam *tree* dihubungkan dengan sebuah aturan, di mana premis terdiri atas sekumpulan *node-node* yang ditemui, dan kesimpulan dari aturam terdiri atas kelas yang terhubung dengan *leaf* dari *path* (Romansyah, 2009).



Gambar 2.2. Konsep Dasar Pohon Keputusan

2.2. Pohon Keputusan ID3

ISBN: 978-979—3649-96-2 Unisbank Semarang, 28 Juli 2016

Algoritma ID3 atau *Iterative Dichotomiser* 3 (ID3) merupakan sebuah metode yang digunakan untuk membuat pohon keputusan yang telah dikembangkan oleh J. Ross Quinlan sejak tahun 1986. Algoritma pada metode ini menggunakan konsep dari *entropy* informasi. Algoritma ini melakukan pencarian secara rakus/menyel*)* pada semua kemungkinan pohon keputusan Anyanwu,. Secara ringkas, langkah kerja Algoritma ID3 dapat digambarkan sebagai berikut (Larose, 2009) :

a. Hitung *Entropy* dan *Information gain* dari setiap atribut dengan menggunakan rumus:

$$Entropy(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Dimana:

S = ruang (data) sample yang digunakan untuk training.

P_+ = jumlah yang bersolusi positif (mendukung) pada data sample untuk kriteria tertentu.

P_- = jumlah yang bersolusi negatif (tidak mendukung) pada data sample untuk kriteria tertentu.

$$Gain(S, A) = Entropy(S) - \sum_{v \in nilai(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Dimana:

S = ruang (data) sample yang digunakan untuk training.

A = atribut.

V = suatu nilai yang mungkin untuk atribut A.

Nilai(A) = himpunan yang mungkin untuk atribut A.

$|S_v|$ = jumlah sample untuk nilai V.

|S| = jumlah seluruh sample data.

Entropy(S_v) = *entropy* untuk sample-sample yang memiliki nilai V.

b. Bentuk simpul yang berisi atribut tersebut.

c. Ulangi proses perhitungan *information gain* yang akan terus dilaksanakan sampai semua data telah termasuk dalam kelas yang sama. Atribut yang telah dipilih tidak diikutkan lagi dalam perhitungan nilai *information gain.*

2.3. **Pohon Keputusan C4.5**

Algoritma *Classification version* 4.5 atau biasa disebut C4.5 adalah pengembangan dari algoritma ID3. Oleh karena pengembangan tersebut, algoritma C4.5 mempunyai prinsip dasar kerja yang sama dengan algoritma ID3. Perbedaan utama C4.5 dari ID3 adalah:

- C4.5 dapat menangani atribut kontinyu dan diskrit.
- C4.5 dapat menangani *training data* dengan *missing value.*
- Hasil pohon keputusan C4.5 akan dipangkas setelah dibentuk.
- Pemilihan atribut yang dilakukan dengan menggunakan *Gain Ratio.*

Information gain pada ID3 lebih mengutamakan pengujian yang menghasilkan banyak keluaran. Dengan kata lain, atribut yang memiliki banyak nilailah yang dipilih sebagai *splitting* atribut. Sebagai contoh, pembagian terhadap atribut yang berfungsi sebagai *unique identifier*, seperti product_ID¸ akan menghasilkan keluaran dalam jumlah yang banyak, di mana setiap keluaran hanya terdiri dari satu *tuple*. Partisi semacam ini tentu saja bersifat *pure*, sehingga informasi yang dibutuhkan untuk mengklasifikasi D berdasarkan partisi seperti ini adalah sebesar Infoproduct_ID(D) = 0. Sebagai akibatnya, *information gain* yang dimiliki atribut product_ID menjadi maksimal. Padahal, jelas sekali terlihat bahwa partisi semacam ini

ISBN: 978-979—3649-96-2 *Unisbank Semarang, 28 Juli 2016*

tidaklah berguna. Karena itu algoritma C4.5 yang merupakan suksesor dari ID3 menggunakan *gain ratio* untuk memperbaiki *information gain*, dengan rumus *gain ratio* (Larose, 2009) :

$$Gain\,Ratio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)}$$

Dimana:
S = ruang (data) sample yang digunakan untuk training.
A = atribut.
Gain(S,A) = *information gain* pada atribut A
SplitInfo(S,A) = *split information* pada atribut A

Atribut dengan nilai *Gain Ratio* tertinggi dipilih sebagai atribut test untuk simpul. Dengan *gain* adalah *information gain.* Pendekatan ini menerapkan normalisasi pada *information gain* dengan menggunakan apa yang disebut sebagai *split information*. *SplitInfo* menyatakan *entropy* atau informasi potensial dengan rumus:

$$SplitInfo(S,A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S}$$

Dimana:
S = ruang (data) sample yang digunakan untuk training.
A = atribut.
S_i = jumlah sample untuk atribut i

2.4. ***Reduced Error Prunning (REP)***

Reduced Error Pruning merupakan salah satu algoritma *postpruning.* Algoritma ini membagi data menjadi dua, yaitu *training data* dan *test data. Training data* adalah data yang digunakan untuk membentuk pohon keputusan, sedangkan *test data* digunakan untuk menghitung nilai *error rate* pada pohon setelah dipangkas. Cara kerja REP adalah dengan memangkas *internal node* yang dimulai dari *internal node* paling bawah ke atas. Pemangkasan dilakukan dengan cara mengganti atribut dengan *leaf node* yang memiliki kelas yang dominan muncul. Setelah itu *test data* diproses menggunakan *rule* hasil pemangkasan, kemudian dihitung nilai *error rate*nya. *Test data* juga diproses dengan rule awal, yaitu rule yang terbentuk sebelum pohon dipangkas, kemudian dihitung nilai *error rate*nya. Apabila nilai *error rate* yang dihasilkan dari pemangkasan pohon lebih kecil, maka pemangkasan dilakukan.

2.5. ***Pre Prunning***

Prepruning yaitu menghentikan pembangunan suatu *subtree* lebih awal, yaitu dengan memutuskan untuk tidak lebih jauh mempartisi *data training*. Rumus *prepruning* (Larose, 2009)

$$e = \frac{r + \frac{z^2}{2n} + z\sqrt{\frac{r}{n} - \frac{r^2}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

Dimana:
r = nilai perbandingan *error rate*
n = total *sample*
$z = \Phi^{-1}(c)$

ISBN: 978-979—3649-96-2 Unisbank Semarang, 28 Juli 2016

c = *confidence level*

3. METODE PENELITIAN

3.1. Pendefinisian Masalah

Tahap ini merupakan tahapan bagaimana suatu permasalahan dirumuskan berdasarkan latarbelakangnya.

3.2. Studi Literatur dan review jurnal

Dukungan teori dan bahan – bahan bacaan, jurnal atau paper mengenai rekayasa perangkat lunak, metode klasifikasi id3 dan c 4.5., *decission support system* (DSS).

3.3. Survey, pengumpulan data dan informasi

Tahap ini bertujuan untuk mengetahui dan melihat secara langsung dan lebih mendetail permasalahan yang akan diteliti, sehingga diperoleh data–data atau informasi yang diperlukan.

Data yang digunakan akan dipecah menjadi beberapa bagian, tergantung dari masing-masing algoritma. ID3 akan memecah data menjadi 2, yaitu:

a. Data *training*: digunakan untuk membentuk pohon keputusan.
b. Data *testing*: digunakan untuk ujicoba pada pohon yang telah dibentuk guna menghitung nilai error rate.

Sedangkan C4.5. akan memecah menjadi 3, yaitu:

a. Data *trainning*: digunakan untuk membentuk pohon keputusan.
b. Data *testing*: digunakan untuk ujicoba pada pohon yang telah dibentuk guna menghitung nilai error rate.
c. Data *testt pruning*: digunakan untuk mengetes akurasi pada pohon yang telah dibentuk guna proses pemangkasan pohon.

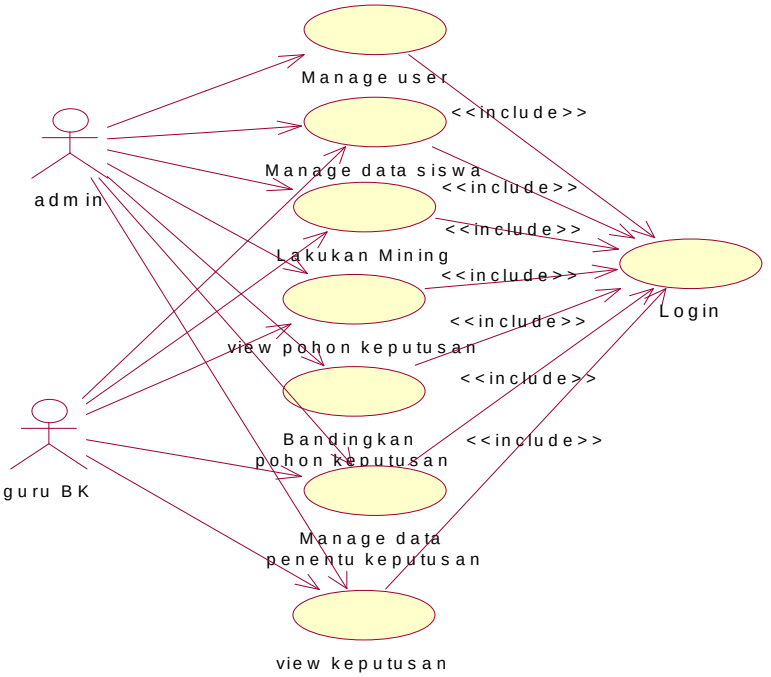
3.4. Analisis kebutuhan sistem

Analisis Kebutuhan sistem adalah gambaran kebutuhan sistem yang akan dibangun secara keseluruhan. Untuk mengidentifikasi gambaran sistem secara keseluruhan adalah dengan melakukan pengamatan (observasi), kemudian mengidentifikasi kebutuhan perangkat lunak yg dibutuhkan untuk Aplikasi penjurusan SLTA.

3.5. Perancangan sistem

Tahap ini adalah tahap dimana dibuat suatu rancangan untuk membangun aplikasi, mulai dari fitur-fitur atau konten, rancangan *user interface*, *Flowchart* sistem dan *flowcha*rt metode, *desain usecase*, *Activity diagram*, desain *data base* (conceptual Data Model / CDM dan Physical Data Model / PDM).

ISBN: 978-979—3649-96-2 *Unisbank Semarang, 28 Juli 2016*



Gambar 3.1. Model *Use Case Diagram Penjurusan SLTA*

3.6. Uji Coba dan Implementasi Sistem

Pada Tahap ini, akan dilakukan terlebih dahulu pengecekan seluruh konten dan prosedur pada program klasifikasi dan di kembangkan menjadi suatu aplikasi perangkat lunak (software klasifikasi penjurusan SLTA).

4. HASIL DAN PEMBAHASAN

4.1. Data yang Digunakan

Data merupakan data kategorikal dan tidak ada *mising value* pada data. Jumlah data yang digunakan sebanyak 200 data. Dalam implementasinya, data dipecah menjadi beberapa bagian, tergantung dari masing-masing algoritma. ID3 dan C4.5 akan memecah data menjadi 2, yaitu:

Proses Mining ID3

Dalam proses mining ID3, proses yang dilakukan adalah sebagai berikut:

1. Hitung frekuensi kemunculan masing-masing nilai atribut pada data survey.
2. Hitung nilai *Entropy* dari masing-masing nilai atribut.
3. Hitung nilai *Information Gain* dengan menggunakan nilai *Entropy* yang telah dihitung sebelumnya.
4. Ambil nilai *Information Gain* terbesar dan jadikan simpul akar.
5. Hilangkan atribut yang dipilih sebelumnya dan ulangi perhitungan nilai *Entropy* dan *Information Gain* dengan memilih *Information Gain* terbesar dan dijadikan simpul *internal* pohon.
6. Ulangi perhitungan tersebut hingga semua atribut pohon memiliki kelas.

ISBN: 978-979—3649-96-2 Unisbank Semarang, 28 Juli 2016

7. Jika semua pohon sudah memilik kelas, maka tampilkan pohon keputusan dan *generate* aturan keputusan.

Proses Mining C4.5

Dalam proses mining C4.5, proses yang dilakukan adalah sebagai berikut:

1. Hitung frekuensi kemunculan masing-masing nilai atribut pada data survey.
2. Hitung nilai *Entropy* dari masing-masing nilai atribut.
3. Hitung nilai *Information Gain* dengan menggunakan nilai *Entropy* yang telah dihitung sebelumnya.
4. Hitung nilai *Split Info* dari tiap atribut.
5. Hitung nilai *Gain Ratio* menggunakan nilai *Information Gain* dan *Split Info*.
6. Ambil nilai *Gain Ratio* terbesar dan jadikan simpul akar.
7. Hilangkan atribut yang dipilih sebelumnya dan ulangi perhitungan nilai *Entropy*, *Information Gain*, *Split Info* dan *Gain Ratio* dengan memilih *Gain Ratio* terbesar dan dijadikan simpul *internal* pohon.
8. Ulangi perhitungan tersebut hingga semua atribut pohon memiliki kelas.
9. Jika semua pohon sudah memilik kelas, maka tampilkan pohon keputusan awal dan generate aturan keputusan awal.

4.2. Skenario 1

Skenario 1 digunakan untuk membandingkan algritma ID3 dan C4.5. pre pruning. Pada skenario 1 ini data yang digunakan yaitu 150 data training dan 50 data testing.

Tabel 4.1: Tabel Data Skenario 1

Jumlah Data	ID3		C4.5	
	Train	Test	Train	Test
IPA (140)	102	38	102	38
IPS (60)	48	12	48	12
Jumlah	150	50	150	50

4.3. Analisa Perbandingan Algoritma

Setelah pohon dibentuk, selanjutnya dilakukan perbandingan dengan data yang merupakan data *testing*, data yang digunakan ada 50 data dimana data tersebut dilakukan pengklasifikasian menggunakan *rule* ID3 dan C4.5. yang telah dibentuk. Kemudian kelas yang terbentuk dibandingkan dan dihitung nilai *error rate*nya. Setelah proses klasifikasi, kemudian dihitung kinerja dari masing-masing algoritma yang meliputi akurasi, error rate, precision dan recall. Berikut tabel kinerja perbandingan :

Tabel 4.2 : Kinerja Perbandingan Algoritma

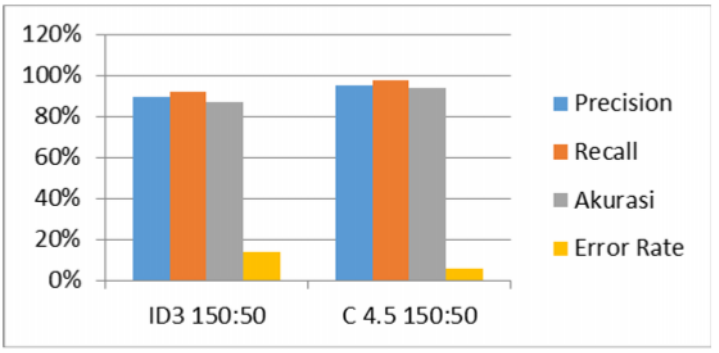
Kinerja	Skenario 1		Skenario 2		
	ID3 150:50	C 4.5 150:50	ID3 125:75	C4.5 125:75	C4.5 Post Pruning 125:75
Akurasi	87%	97%	93%	94%	96%

ISBN: 978-979—3649-96-2 Unisbank Semarang, 28 Juli 2016

Error Rate	14%	6%	8%	7%	5%
Precision	89,74 %	94,87%	93,42%	93,51%	94,74%
Recall	92,11 %	97,37%	97,26%	98,63%	98,63%

Pada Skenario 1 terdapat penilaian kinerja algoritma ID3 dan C4.5. Penilaian kinerja diperoleh dari hasil klasifikasi rule algoritma dengan data testing. Perbandingan skenario 1 ini digunakan untuk membandingkan kinerja dari kedua algoritma, guna mengetahui algoritma mana yang paling bagus kinerjanya. Dari Hasil penilaian kinerja diketahui algoritma C 4.5 memiliki akurasi yang lebih baik dari pada ID3. Ini terlihat dari nilai akurasi C4.5 sebesar 96% sedangkan ID3 sebesar 87%.

Perbandingan pada skenario 1 dari kedua algoritma dapat digambarkan pada grafik berikut:



Gambar 4.1: Grafik Skenario 1

5. KESIMPULAN

Adapun kesimpulan dari

1. Dari pengukuran kinerja kedua algoritma yang telah dilakukan, dapat disimpulkan algoritma C4.5 memiliki kinerja (*precision*, *recall*, dan *accuracy*) yang lebih baik dibandingkan algoritma ID3
2. Metode post pruning merupakan metode pruning yang lebih baik dari pada pre pruning, hal ini dapat dilihat pada partisi data dimana post pruning memiliki nilai yang lebih baik dari pada pre pruning

DAFTAR PUSTAKA

ISBN: 978-979—3649-96-2

Anyanwu, Matthew N., and Shiva Sajjan G. 2010, Comparative Analysis of Serial Decision Tree Classification Algorithms. *International Journal of Computer Science and Security, (IJCSS) Volume (3) : Issue (3)*. 1: 2-4.

Hardikar S, Shrivastava A, Choudhary V. 2012, *Comparison between ID3 and C4.5 in Contrast to IDS*. VSRD-IJCSIT. Vol. 2 (7). 659-667.

Kumar, S. Anupama, dan M.N, Vijayalakshmi. 2011, Efficiency of Decision Trees in Predicting Student's Academic Performance. *Computer Science & Information Technology (CS & IT) DOI: 10.5121/csit.2011.1230*. 1: 5-8.

Larose, Daniel T. 2005, *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey : John Willey & Sons, Inc.

Nugroho, Fanuel., Kristanto, Harianto.,2007, dan Oslan, Yetli. Validitas Suatu Alamat menggunakan Pohon keputusan dengan Algoritma ID3. *Jurnal Informatika, Volume 3 Nomor 2*

Sunjana. 2010, Klasifikasi Data Nasabah Sebuah Asuransi menggunakan Algoritma C4.5. *Seminar Nasional Aplikasi Teknologi Informasi ((SNATI 2010) ISSN: 1907-5022*. 1: 1-2.

Yusuf, Y.W. 2007,*Perbandingan Performansi Algoritma Decision Tree C5.0, CART, Dan CHAID: Kasus Prediksi Status Resiko Kredit di Bank X*. Seminar Nasional Aplikasi Teknologi Informasi 2007**.** ISSN:1907-5022. B59-B62.

Yeni kust iyahningsih, 2014, Penentuan pemberian Kredit UKM menggunakan metode ID3, Semnastek, UGM, Yogyakarta

Romansyah, F., Sitanggang I. S., dan Nurdiati, S.2009, Fuzzy Decision Tree dengan Algoritma ID3 pada Data Diabetes. *Internetworking Indonesia Journal Vol. 1/No. 2*

Sharma, Aman K., dan Sahni, Suruchi. 2012, A Comparative Study of Classification Algorithms for Spam Email Data Analysis. *International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397*. 1: 2-5. 2011.