PERBANDINGAN *DECISION TREE* PADA ALGORITMA C 4.5 DAN ID3 DALAM PENGKLASIFIKASIAN INDEKS PRESTASI MAHASISWA (Studi Kasus: Fasilkom Universitas Singaperbangsa Karawang)

Jejen Jaenudin, Prabowo Pudjo Widodo
Unversitas Ibn Khaldun Bogor
Jln. K.H Sholeh Iskandar Km. 2 Bogor
zen@ft.uika-bogor.ac.id, prabowo_pw@yahoo.com

Abstract- This study applied a data mining model of student's grade point classification at the department of information technology on computer science faculty singaperbangsa university karawang. The objective of the study is to comprehend the description of rule models which are obtained to produce a decision based on "satisfied and dissatisfied" predicates. The decision tree algorithm on this study are the C 4.5 and ID3 algorithms. For data analysis, this study uses supporting software of RapidMiner 5.1. In designing data mining process, this research uses Cross-Industry Standard Process for Data Mining (CRISP-DM) model. The resulting output is C 4.5 decision tree algorithm might support the computer science faculty of singaperbangsa university karawang on decision-making in teaching and learning process as the basic reference for future improvements.

Keywords: *Data Mining, Decision Tree, Precision, Recall, Accuracy*

I. PENDAHULUAN

A. Latar Belakang

Pada saat ini, informasi telah menjadi kebutuhan pokok bagi kehidupan manusia. Hampir di setiap tempat, informasi selalu berguna mendukung aktivitas manusia, demikian halnya dengan usaha-usaha manusia untuk melakukan pengelolaan sumber daya organisasi. Pengelolaan suatu obyek seperti sumber daya organisasi memerlukan informasi untuk suatu analisis perencanaan, pengambilan keputusan, dan pelaksanaan. Ketidaktersediaan informasi akan menyebabkan suatu perencanaan menjadi tidak fokus pada obyek yang ditangani, sehingga akan menyebabkan suatu keputusan atau kebijakan yang diambil tidak dapat menyelesaikan masalah-masalah yang ada. Guna mendapatkan informasi yang akurat dan mudah dalam pengaksesannya, maka informasi dapat dikemas dalam sebuah sistem yang berbasis komputer.

Data mining semakin hari semakin penting, karena semakin dibutuhkan dalam menyelesaikan masalah-masalah nyata baik di dalam dunia *sains*, rekayasa, industri, pemerintahan maupun pendidikan. Apalagi dengan perkembangan teknologi pengumpulan

data saat ini, di mana jumlah data yang dikumpulkan per unit waktu semakin bertambah banyak dengan kecepatan yang berlipat. Hal ini menambah volume data yang tersimpan dan harus diolah semakin membesar. Peningkatan volume data yang besar memerlukan metode yang bisa bekerja cepat dan terotomatisasi untuk mengolah dan mengambil kesimpulan dari data tersebut.

Dalam penelitian ini mengaplikasikan teknik *data mining* dengan algoritma C 4.5 dan algoritma *iterative dichotomiser* 3 (ID3) dalam membuat model aturan *decision tree* untuk mendukung peningkatan indeks prestasi mahasiswa pada Fasilkom UNSIKA. Model aturan tersebut diperoleh dari pengklasifikasian indeks prestasi mahasiswa berdasarkan predikat puas dan tidak puas. Kemudian membandingkan kedua algoritma tersebut dan memilih mana di antara kedua yang lebih unggul kinerjanya serta diharapkan dapat memberikan kontribusi terhadap Fasilkom UNSIKA sebagai acuan dalam proses kegiatan belajar dan mengajar terhadap bidang ilmu komputer yang akan diambil pada angkatan berikutnya.

B. Tujuan

Tujuan yang ingin dicapai dengan adanya penelitian ini adalah:

1. Untuk mendapatkan sebuah model aturan dalam mendukung peningkatan indeks prestasi mahasiswa pada Fasilkom UNSIKA.
2. Untuk membantu Fasilkom UNSIKA dalam mengambil tindakan prepentif terhadap atribut dari indeks prestasi mahasiswa yang tidak memuaskan.

II. METODOLOGI

Seperti yang telah dijelaskan sebelumnya, bahwa metode yang penulis gunakan dalam penelitian ini menggunakan model *Cross-Standard Industry for Data Mining* (CRISP-DM). Dalam CRISP-DM ini, sebuah proyek *data mining* memiliki siklus hidup yang terbagi dalam enam fase, yaitu:

1. Fase Pemahaman Bisnis

Dari laporan panitia penerimaan mahasiswa baru, Fasilkom UNSIKA dari tahun ke tahun minat calon mahasiswa yang mendaftar terhadap program studi yang ada di Fasilkom selalu bertambah bahkan dari beberapa tahun ini, fakultas ini merupakan satu di antara fakultas-fakultas lain yang paling banyak diminati.

Dengan semakin banyaknya minat calon mahasiswa maka perlu adanya pembenahan KBM yang lebih bagus lagi, maka Fasikom UNSIKA mempunyai

tujuan dan tekad untuk menghasilkan lulusan yang mampu menggunakan komputer dalam proses rekayasa, menguasai teknik dan metode penyelesaian masalah dengan bantuan komputer serta mampu mengembangkan kegiatan penelitian dan mampu meneruskan studi lanjutan dalam hal bidang keilmuan dan teknologi komputer. Oleh karena itu, untuk mempersiapkan itu semua perlu kiranya untuk melihat kemampuan belajar dan kemajuan mahasiswa yang dilakukan pada evaluasi hasil belajar pada setiap semesternya yang kemudian dinyatakan dengan indeks prestasi mahasiswa.

Berdasarkan data yang diperoleh penulis dari bagian akademik Fasilkom, sampai saat ini belum diketahui kriteria model *decision tree* untuk algoritma dalam melakukan pengklasifikasian indeks prestasi mahasiswa untuk mendapatkan hasil aturan yang akurat. Oleh karena itu, maka dalam penelitian ini akan dilakukan pengujian model serta perbandingan antara algoritma *decision tree* tersebut.

2. Fase Pemahaman Data

Untuk menentukan klasifikasi indeks prestasi mahasiswa, enam atribut *predictor* dan satu atribut *class*. Atribut-atribut yang menjadi parameter terlihat pada Tabel berikut ini.

Tabel 1: Atribut dan Nilai Kategori

No	Atribut	Nilai
	Algoritma & Pemrograman	
1	Lanjutan	< 70 >=70
2	Metode Numerik	< 70 >=70
3	Organisasi Komputer	< 70 >=70
4	Statistik 1	< 70 >=70
5	Struktur Data	< 70 >=70
	Teknik Pemrograman	
6	Terstruktur	< 70 >=70

3. Fase Pengolahan Data

Data yang diperoleh untuk penelitian ini sebanyak 485 data *record* mahasiswa. Untuk mendapatkan data yang berkualitas, beberapa teknik *preprocessing* digunakan[6], yaitu:

1. *Data validation*, untuk mengidentifikasi dan menghapus data yang ganjil (*outlier/ noise*), data yang tidak konsisten dan data yang tidak lengkap (*missing value*).

2. *Data integration and transformation*, untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal.
3. *Data size reduction and dicretization*, untuk memperoleh data set dengan jumlah atribut dan *record* yang lebih sedikit tetapi bersifat informatif. Di dalam data *training* yang digunakan dalam penelitian ini, dilakukan seleksi atribut dan penghapusan data duplikasi menggunakan *software* RapidMiner 5.1.

Setelah dilakukan *preprocessing* data yang didapat dari data mahasiswa sebanyak 485 *record* direduksi dengan menghilangkan duplikasi menjadi 211 *record*.

4. Fase Pemodelan

Tahap ini juga dapat disebut tahap *learning* karena pada tahap ini, data *training* diklasifikasikan oleh model dan kemudian menghasilkan sejumlah aturan. Pada penelitian ini, pembuatan model menggunakan dua algoritma yaitu algoritma C 4.5 dan ID3.

5. Fase Evaluasi

Pada tahap ini dilakukan pengujian terhadap model-model yang dibandingkan untuk mendapatkan informasi model yang paling akurat. Evaluasi dan validasi menggunakan *confusion matrix* dan kurva ROC.

6. Fase Penyebaran

Setelah pembentukan model dilakukan analisa dan pengukuran pada tahap sebelumnya, selanjutnya pada tahap ini diterapkan model yang paling akurat untuk pengklasifikasian indeks prestasi mahasiswa.

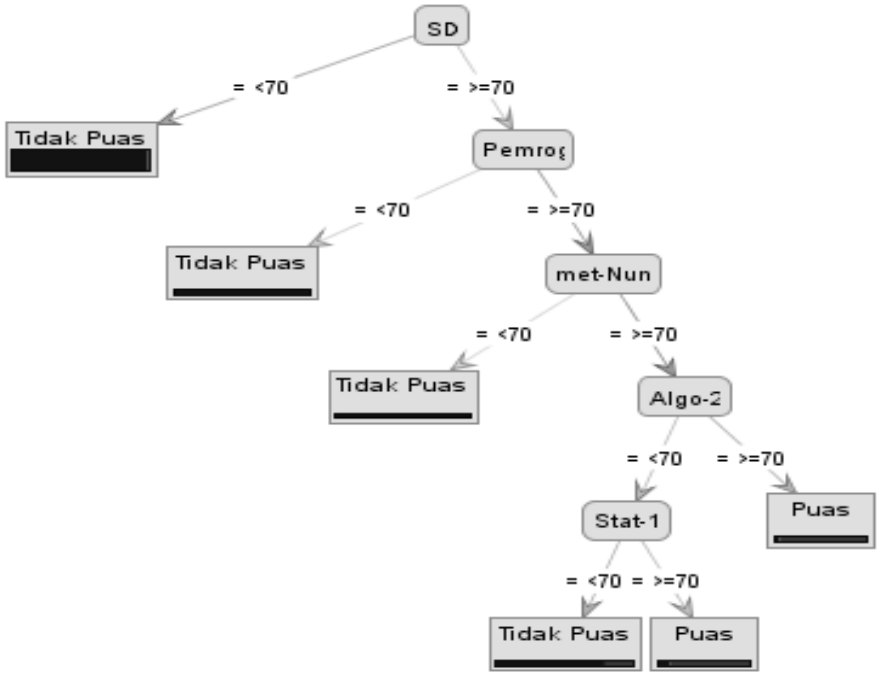
III. HASIL DAN BAHASAN

A. Hasil

Pada langkah awal dari proses perhitungan algoritma C 4.5 ini adalah menentukan terlebih dahulu atribut sebagai akarnya yaitu Algoritma & Pemrograman 2, Metode Numerik, Organisasi Komputer, Statistik 1, Struktur Data dan Pemrograman Terstruktur. Atribut yang terpilih adalah yang mempunyai nilai *gain* tertinggi dari nilai *gain-gain* atribut lainnya, kemudian dijadikan akar dari pohon.

Dari hasil yang didapat untuk data *training* pada lampiran 3, dengan menggunakan *tools* RapidMiner 5.1

model pohon keputusan yang didapatkan dapat dilihat sebagai berikut.



Gambar 1 Pohon Keputusan C 4.5 dengan RapidMiner 5.1

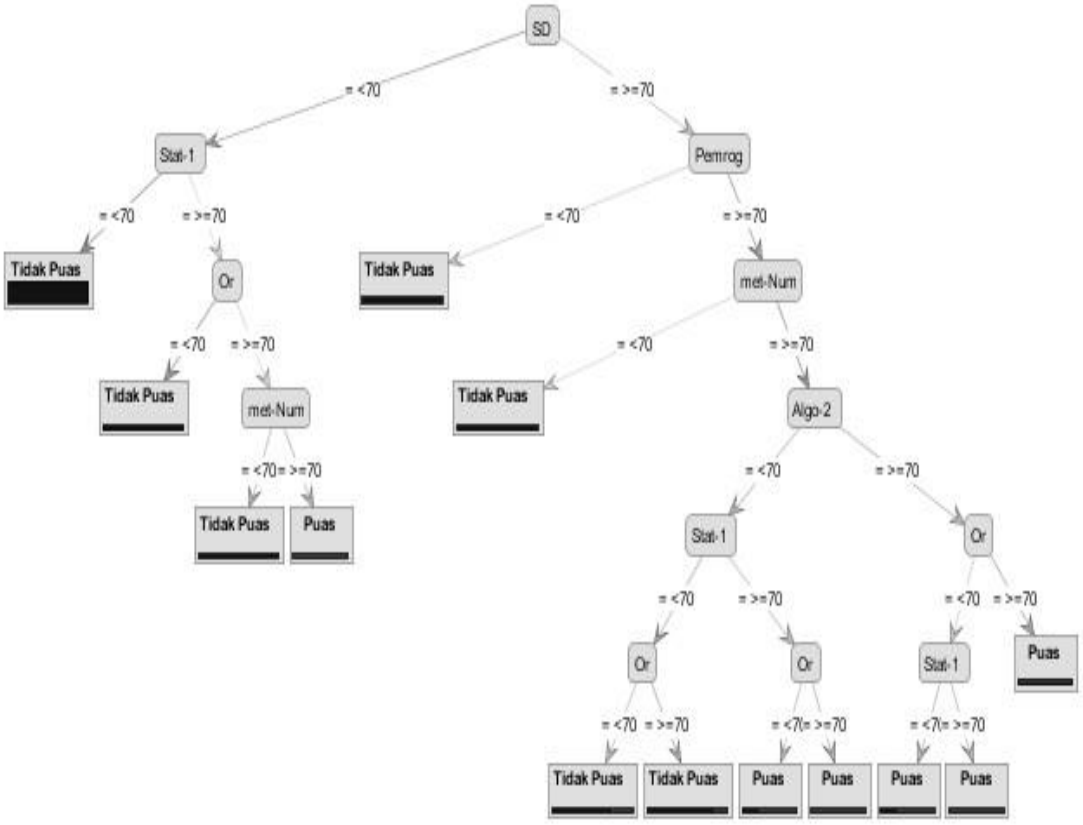
Berdasarkan pohon keputusan yang terlihat pada gambar 4.6 di atas, dapat dibentuk aturan-aturan untuk model pohon keputusan dari algoritma C 4.5 sebagai berikut:

1. R1: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik >=70 AND Algoritma 2 >=70 THEN Puas
2. R2: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik >=70 AND Algoritma 2 <70 AND Statistik 1 >=70 THEN Puas
3. R3: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik >=70 AND Algoritma 2 <70 AND Statistik 1 <70 THEN Tidak Puas
4. R4: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik <70 THEN Tidak Puas
5. R5: If Struktur Data >=70 AND Pemrograman Terstruktur <70 THEN Tidak Puas
6. R6: If Struktur Data <70 THEN Tidak Puas

Pada langkah selanjutnya dalam membentuk suatu model pohon keputusan dengan menggunakan model algoritma ID3 adalah seperti halnya pada langkah dalam penentuan model algoritma C 4.5.

Tahap awal dalam pembuatan suatu model dalam algoritma ID3 ditentukan terlebih dahulu atribut sebagai akarnya yaitu Algoritma & Pemrograman 2, Metode Numerik, Organisasi Komputer, Statistik 1, Struktur Data dan Pemrograman Terstruktur. Atribut yang terpilih adalah yang mempunyai nilai *gain* tertinggi dari nilai *gain-gain* atribut yang lainnya, kemudian dijadikan akar dari pohon.

Dari hasil yang didapat untuk data *training* pada lampiran 3 dengan menggunakan *tools* RapidMiner 5.1, model pohon keputusan untuk algoritma ID3 yang didapatkan dapat dilihat sebagai berikut.



Gambar 2 Pohon Keputusan ID3 dengan RapidMiner 5.1

Berdasarkan pohon keputusan yang terlihat pada gambar di atas, dapat dibentuk aturan-aturan untuk model pohon keputusan pada algoritma ID3 adalah sebagai berikut:

1. R1: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik >=70 AND Algoritma 2 >=70 AND Organisasi Komputer >=70 THEN Puas
2. R2: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik >=70 AND Algoritma 2 >=70 AND Organisasi Komputer <70 AND Statistik 1 >=70 THEN Puas
3. R3: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik >=70 AND Algoritma 2 >=70 AND Organisasi Komputer <70 AND Statistik 1 < 70 THEN Puas
4. R4: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik >=70 AND Algoritma 2 <70 AND Statistik 1 >=70 AND Organisasi Komputer >=70 THEN Puas
5. R5: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik >=70 AND Algoritma 2 <70 AND Statistik 1 >=70 AND Organisasi Komputer <70 THEN Puas
6. R6: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik >=70 AND Algoritma 2 <70 AND Statistik 1 <70 AND Organisasi Komputer >=70 THEN Tidak Puas
7. R7: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik >=70 AND Algoritma 2 <70 AND Statistik 1 <70 AND Organisasi Komputer <70 THEN Tidak Puas
8. R8: If Struktur Data >=70 AND Pemrograman Terstruktur >=70 AND Metode Numerik <70 THEN Tidak Puas
9. R9: If Struktur Data >=70 AND Pemrograman Terstruktur <70 THEN Tidak Puas
10. R10: If Struktur Data <70 AND Statistik 1 >=70 AND Organisasi Komputer >=70 AND Metode Numerik >=70 THEN Puas
11. R11: If Struktur Data <70 AND Statistik 1 >=70 AND Organisasi Komputer >=70 AND Metode Numerik <70 THEN Tidak Puas
12. R12: If Struktur Data <70 AND Statistik 1 >=70 AND Organisasi Komputer <70 THEN Tidak Puas

13. R13: If Struktur Data <70 AND Statistik 1 <70 THEN Tidak Puas

Proses *data mining* yang penulis gunakan dalam penelitian ini dilakukan dengan bantuan *software data mining* yaitu RapidMiner 5.1. Algoritma yang akan diujikan dalam penulisan ini adalah algoritma ID3 dan C4.5 yang berada pada modul *classify*. Kemudian dilakukan komparasi di antara keduanya dan mengukur metode mana yang paling tinggi tingkat akurasinya dengan menggunakan *confussion matrix* dan kurva ROC/AUC (*Area Under Cover*).

Hasil dari pengujian model yang telah dikembangkan akan diuji tingkat keakuratannya dengan memasukkan sejumlah data uji (data *testing*) ke dalam model. Menurut Han dan Kamber (2006), untuk mengukur kekakuratan model dengan baik, data uji seharusnya bukan data yang berasal dari data *training*. Data uji yang penulis gunakan dalam penelitian ini berasal dari data indeks prestasi mahasiswa Fasilkom UNSIKA tahun 2008-2011. Ada 35 *sampel* yang diambil untuk mendapatkan tingkat akurasi data dari data keseluruhan dengan jumlah total data sebanyak 211 dengan tingkat kesalahan 2,5 % secara acak *(simple random sampling)*.

Berdasarkan hasil pengujian tingkat akurasi dari beberapa Tabel *confusion matrix* di atas, selanjutnya didapatkan hasil perhitungan nilai *accuracy, precision* dan *recall*. Nilai *accuracy, precision* dan *recall* yang telah dihitung untuk algoritma C 4.5 dan ID3 tersebut dapat dilihat pada Tabel berikut.

Tabel 1 Nilai A*ccuracy, Precision*, dan *Recall*

	Data Training		**Data Testing**	
	C 4.5	ID3	C 4.5	ID3
Accuracy	96.67 %	96.11 %	95.00 %	91.67 %
Precision	95.00 %	90.00 %	96.67 %	95.00 %
Recall	85.00 %	85.00 %	96.67 %	86.67 %

Berdasarkan hasil pengujian tingkat akurasi dari beberapa gambar ROC di atas, selanjutnya didapatkan hasil perhitungan nilai ROC. Nilai ROC yang telah dihitung untuk algoritma C 4.5 dan ID3 tersebut dapat dilihat pada Tabel berikut.

Tabel 2: Nilai ROC

Algoritma	Confusion Matrix		ROC	
	Training	Testing	Training	Testing
C 4.5	96.67%	95.00%	0.968	0.950
ID3	96.11%	91.67%	0.930	0.500

Berikut hasil komparasi nilai *accuracy* dan nilai ROC yang ditampilkan pada Tabel di bawah sebagai berikut:

Tabel 3 Komparasi nilai *accuracy* dan ROC

Algoritma	Confusion Matrix		Perbandingan Komparasi
	Training	Testing	
C 4.5	96.67%	95.00%	96.34%
ID3	96.11%	91.67%	95.22%

Dari hasil pengujian di atas, dengan dilakukan evaluasi. Pada Tabel 3 dengan memperhatikan kolom *training* pada akurasi, algoritma C 4.5 memiliki tingkat akurasi yang paling tinggi dengan tingkat akurasi 96.67%, demikian pula pada kolom *testing* algoritma C 4.5 memiliki nilai akurasi paling tinggi yaitu 95.00%.

Selanjutnya berdasarkan kolom ROC pada Tabel 3, pada *training* algoritma C 4.5 memiliki tingkat ROC paling tinggi, yaitu 0.968 dan pada *testing* algoritma C 4.5 juga memiliki tingkat ROC yang paling tinggi, yaitu 0.950 termasuk dalam katagori klasifikasi sangat baik.

Dengan mengunakan perbandingan data *training* dengan data *testing*, yaitu 80 berbanding 20, maka untuk akurasi dapat dilihat dalam Tabel 4 sebagai berikut:

Tabel 4 Perbandingan Akurasi

Data Training		**Data Testing**	
C 4.5	ID3	C 4.5	ID3
0.968	0.930	0.950	0.500

Oleh karena itu, berdasarkan Tabel 4.25 algoritma C 4.5 memiliki tingkat akurasi yang paling tinggi dibandingkan dengan algortima ID3, sehingga baik digunakan untuk pengklasifikasian indeks prestasi mahasiswa dengan *persentase* 96.34%.

B. Implikasi Penelitian

Berdasarkan hasil penelitian yang telah dilakukan ini, diharapkan dapat memberikan inspirasi dan masukan bagi pihak Fasilkom UNSIKA untuk dapat memanfaatkan teknik *data mining*,

khususnya metode klasifikasi yang dapat membantu dalam pengambilan keputusan pada indeks prestasi mahasiswa. Dari hasil evaluasi dari perhitungan kedua algoritma di atas, ternyata algoritma C4.5 terbukti paling akurat dibanding algoritma ID3.

IV. PENUTUP

Dari pengukuran kinerja kedua algoritma yang telah dilakukan berdasarkan jumlah data maka dapat disimpulkan bahwa algoritma C 4.5 memiliki kinerja yang lebih baik. Maka hasil penelitian dari percobaan yang telah dilakukan dapat disimpulkan bahwa algoritma C 4.5 mempunyai kinerja yang lebih baik dibandingkan algoritma ID3.

Berdasarkan hasil penelitian yang diperoleh penulis, kiranya perlu pengembangan *riset* lebih lanjut demi kesempurnaan dan keakurasian data yang telah dikembangkan. Oleh karena itu, untuk melihat tingkat akurasi dari algoritma, akan lebih baik lagi kedua algoritma C 4.5 dan ID3 dibandingkan atau dikomparasi dengan model algoritma lain seperti *Support Vector Machine*, *Naive Bayes*, K-NN, *Linear Regression* ataupun algoritma lainnya.

V. DAFTAR PUSTAKA

[1] Han, J., & Kamber, M. 2006. ”*Data Mining: Consepts, Models, and Techniques”.* Fransisco: Morgan kauffman.

[2] Kadir, “Pengenalan Teknologi Informasi”, Andi Offset, Yogyakarta, 2003.

[3] Kusrini & Luthfi, E. T. 2009. ”*Algoritma Data Mining.* Yogyakarta: Andi Publishing.

[4] Larose, D. T. 2005. ”*Discovering Knowledge in Data”. New Jersey: John Willey & Sons, inc.*

[5] Munir, “Buku Teks Ilmu Komputer Matematika Diskrit”, Informatika, Bandung, 2001.

[6] Vercellis, 2009, Business Intelligence: ”*Data Mining and Optimization for Decision making”*.