Faktor Exacta *11 (3):* 266-274, 2018
p-ISSN: 1979-276X
e- ISSN: 2502-339X
DOI : 10.30998/faktorexacta.v11i3.2777

ALGORITMA KLASIFIKASI DATA MINING UNTUK MEMPREDIKSI SISWA DALAM MEMPEROLEH BANTUAN DANA PENDIDIKAN

SENNA HENDRIAN
_Program Studi Informatika
Universitas Idraprasta PGRI
Jl. Raya Tengah No. 80, Kel. Gedong, Kec. Pasar Rebo, Jakarta Timur 13760
Email: Senna.hendrian@unindra.ac.id

Abstrak: Pendidikan merupakan salah satu komponen kehidupan yang dapat menunjang keberhasilan seseorang menuju kehidupan yang jauh lebih baik lagi. Terutama bagi anak yang ada dalam lingkup usia wajib belajar. akan tetapi tidak semua anak wajib belajar dapat mengikuti pendidikan, karena beberapa faktor penyebab, salah satunya adalah masalah biaya pendidikan. Untuk mengatasi permasalahan yang ada, maka sekolah Bina Bangsa Mandiri menyusun program Bantuan Dana Pendidikan bagi Siswa yang dianggap kurang mampu secara strata ekonomi. Pada penelitian ini penulis menggunakan Algoritma Klasifikasi *Dataminig* yaitu Algoritma C4.5 untuk memprediksi siswa dalam memperoleh bantuan dana pendidikan. Sampel data diambil dari SMA Bina Bangsa Mandiri yang beralamat di Kecamatan Gunungputri Kab. Bogor. Dari hasil pengujian digunakan tes *Cros Validation* dan *Confusion Matrix* dan Kurva ROC. Hasil yang diperoleh untuk nilai *Accuracy* Algoritma C4.5 adalah sebesar **98,80%**, nilai untuk *Precision* sebesar **98,02%**, dan nilai untuk *Sensitivity* atau *Recall* sebesar **99,00%.** Dengan demikian Algoritma C4.5 merupakan algoritma dan teknik terbaik untuk Memprediksi Siswa dalam memperoleh Bantuan Dana Pendidikan.
Kata Kunci : data mining, algoritma klasifikasi, algoritma C4.5, dana pendidikan.

Abstract: *Education is one of the components of life that can support the success of a person towards life that much better again. Especially for the children that are in the scope of the age of compulsory education. but not all children can attend compulsory education, because several factors cause, one of which is the issue of tuition fees. To cope with the existing problems, then a standalone compiled Bina Bangsa School programs Help Fund education for students who are considered less capable in economic strata. In this study, the author uses the classification Algorithm Datamining Algorithm C4.5 to predict students in obtaining the help of the Education Fund. Sample data are drawn from the Upper secondary school (HIGH SCHOOL) Self-sustaining Bina Bangsa (BBM) that located in Kecamatan Gunungputri Kab. Bogor. From the results of testing and Validation of tests used Cros Confusion Matrix and ROC Curves. The results obtained for the value of Accuracy Algorithm C 4.5 is 98.80%, a value for the Precision of 98.02%, and the value for Sensitivity or Recall of 99.00%. Thus the algorithm C 4.5 is the best techniques and algorithms to predict Students in obtaining the help of the Education Fund.*
Keywords : *data mining, algorithms of classification, algorithm C 4.5, the education fund*

PENDAHULUAN

Ilmu pengetahuan seseorang dapat diperoleh melalui pendidikan di sekolah, karena dengan bersekolah kita akan mampu mewujudkan keberhasilan dan kesuksesan dalam kehidupan. Namun empirik dilapangan menunjukan bahwa ekonomi yang terbatas bagi sebagian orang tua menjadi faktor penghambat dalam mewujudkan kesuksesan anaknya, sehingga tidak semua anak usia wajib belajar dapat mengikuti pendidikan di sekolah. Untuk mengatasi permasalahan yang ada, yaitu melalui program beasiswa pendidikan. Yayasan Bina Bangsa Mandiri merupakan salah satu lembaga pendidikan dengan mempunyai tiga Sekolah, yaitu SMP, SMA dan SMK Bina Bangsa Mandiri, dimana sekolah ini mempunyai program Bantuan Dana Pendidikan bagi Siswa yang

Faktor Exacta *11 (3):* 266-274, 2018
p-ISSN: 1979-276X
e- ISSN: 2502-339X
DOI : 10.30998/faktorexacta.v11i3.2777

dianggap kurang mampu secara strata ekonomi. Namun sudah tentu syarat dan ketentuan berlaku, dan semua kriteria pemilihan penentuan siswa memperoleh bantuan dana pendidikan itu dilakukan berdasarkan data siswa yang ada, kemudian dianalisis secara manual, namun terkadang hasil yang diperoleh tidak sesuai. Dengan demikian pada penelitian ini penulis menggunakan Algoritma Klasifikasi *Dataminig* yaitu Algoritma C4.5 untuk memprediksi siswa dalam memperoleh bantuan dana pendidikan. Sampel data diambil dari Sekolah Menengah Atas (SMA) Bina Bangsa Mandiri yang beralamat di Kecamatan Gunungputri Kabupaten Bogor.

Data mining telah banyak menarik perhatian di masyarakat dalam beberapa tahun ini, karena mampu mengubah data yang luas dan jumlah yang besar menjadi informasi yang berguna dan pengetahuan. Informasi dan pengetahuan yang diperoleh dapat digunakan untuk mengaplikasikan seperti analisis pasar, deteksi penipuan, dan retensi pelanggan, untuk pengendalian produksi dan ilmu pengetahuan eksplorasi (Han & Kember, 2007).

Sementara menurut Turban, E. dkk 2005, data mining merupakan sebuah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar.

Klasifikasi adalah tipe analisis data yang dapat membantu orang menentukan kelas label dari sampel yang ingin di klasifikasi. Klasifikasi merupakan Metode *supervised learning*, metode yang mencoba menemukan hubungan antara atribut masukan dan atribut target. Tujuan klasifikasi untuk meningkatkan kehandalan hasil yang diperoleh dari data.

Algoritma C4.5 Salah satu metode klasifikasi yang digunakan. Melibatkan konstruksi pohon keputusan, koleksi node keputusan. Setiap cabang kemudian mengarah ke *node* lain baik keputusan atau ke *node* daun untuk mengakhiri (Larose, 2005). C4.5 adalah algoritma yang mempunyai *input* berupa training *samples* berupa data contoh yang akan digunakan untuk membangun sebuah *tree* yang telah diuji kebenarannya dan *samples* yang merupakan *field - field* data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data. Algoritma dasar dari C4.5 adalah sebagai berikut:

1). Pohon yang dihasilkan berupa pohon terbalik, 2). Pada tahap awal, semua contoh *training* adalah akar. 3). Atribut adalah kategori. 4) Contoh di partisi secara berulang berdasarkan atribut yang dipilih. 5). Atribut tes dipilih dari data *heuristic* atau pengukuran statistik

Tahapan algoritma C4.5 adalah sebagai berikut:

1). Siapkan data training. 2) Pilih atribut sebagai akar

Untuk memilih atribut akar, didasarkan pada nilai *Gain* tertinggi dari atribut-atribut yang ada. Untuk mendapatkan nilai *Gain*, harus ditentukan terlebih dahulu nilai *Entropy*.

Rumus *Entropy* :

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2\ pi$$

S = Himpunan Kasus
n = Jumlah Partisi S
pi = Proporsi dari Si terhadap S

Rumus *Gain* :

$$Gain(S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (S_i)$$

S = Himpunan Kasus
A = Atribut
n = Jumlah Partisi Atribut
|Si| = Jumlah Kasus pada partisi ke-i
|S| = Jumlah Kasus dalam S

Faktor Exacta *11 (3):* 266-274, 2018
p-ISSN: 1979-276X
e- ISSN: 2502-339X
DOI : 10.30998/faktorexacta.v11i3.2777

Decission Tree atau Pohon Keputusan adalah struktur sederhana yang dapat digunakan sebagai pengklasifikasi. Referensi penting dalam pengerjaan aslinya adalah *Classification and Regression Tree* oleh Breiman et al.

Pada pohon keputusan, masing-masing node internal (*non-leaf*) merepresentasikan sebuah variabel atribut (atribut prediksi atau fitur) dan masing-masing cabang merepresentasikan satu keadaan dari variabel ini. Masing-masing dari tiga daun (*leaf*) menspesifikasikan nilai yang diharapkan dari kelas variabel (variabel yang akan di prediksi). Aspek penting dari prosedur untuk membangun pohon keputusan adalah pemisahan kriteria (*split criterion*) termasuk kriteria untuk membuat cabang dan kriteria terakhir (*stop criterion*), kriteria yang digunakan untuk menghentikan pencabangan. Pohon keputusan dibuat menggunakan himpunan dari data yang digunakan sebagai data pembelajaran (*training dataset*). Himpunan yang berbeda yang disebut *test dataset* digunakan untuk melakukan pengujian untuk mengecek model. Pohon keputusan menawarkan banyak keuntungan, antara lain:

1. Fleksibilitas untuk berbagai tugas *data mining*, seperti klasifikasi, regresi, clustering dan seleksi fitur.
2. Cukup jelas dan mudah diikuti (ketika dipadatkan).
3. Fleksibilitas dalam menangani berbagai input data: nominal, numerik dan tekstual.
4. Adaptasi di dataset pengolahan yang mungkin memiliki kesalahan atau nilai-nilai yang hilang.
5. Kinerja prediktif tinggi untuk upaya komputasi yang relatif kecil.
6. Tersedia dalam berbagai paket *data mining* melalui berbagai *platform*
7. Berguna untuk dataset besar (dalam kerangka *ensemble*).

Cross Validation merupakan salah satu teknik untuk menilai/memvalidasi keakuratan sebuah model yang dibangun berdasarkan dataset tertentu. Data yang digunakan dalam proses pembangunan model disebut data latih/*training*, sedangkan data yang akan digunakan untuk memvalidasi model disebut sebagai data *test*.

Tabel 1. *Confusion Matrix*

Kelas	**Prediksi *Yes***	**Prediksi *No***	**Total**
Aktual *Yes*	*True Positive* (TP)	*False Negative* (FN)	*Positive* (P)
Aktual *No*	*False Positive* (FP)	*True Negative* (TN)	*Negative* (N)
Total	P’	N’	P+N

Rumus Mencari Akurasi adalah :

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Rumus Mencari *Sentitivity* atau *Recall* adalah :

$$sensitivity = \frac{TP}{TP + FN}$$

Rumus Mencari Nilai *Precision* adalah :

$$precision = \frac{TP}{TP + FP}$$

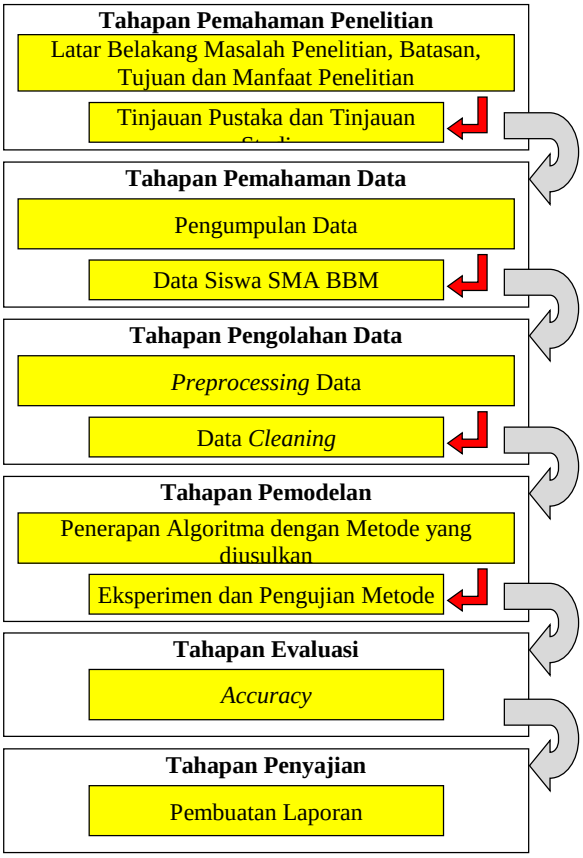
Kurva *ROC (Receiver Operating Characteristic)* menunjukan akurasi dan membandingkan klasifikasi secara visual. *ROC* mengekspresikan *confusion matrix. ROC* adalah grafik dua dimensi dengan *false positives* sebagai garis horizontal dan *true positives* untuk mengukur perbedaaan performasi metode yang digunakan. *ROC Curve* adalah cara lain untuk menguji kinerja pengklasifikasian (Gorunescu, 2010). Selanjutnya dari grafik ROC diperoleh nilai *Area Under Curve* (UAC)

Faktor Exacta *11 (3):* 266-274, 2018
p-ISSN: 1979-276X
e- ISSN: 2502-339X
DOI : 10.30998/faktorexacta.v11i3.2777

Rapid Miner merupakan perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari Institute of Technology Blanchardstown dan Raif Klinkenberg dari rapid-i.com dengan tampilan GUI (*Graphical User Interface*) sehingga memudahkan pengguna dalam menggunakan perangkat lunak ini. Perangkat lunak ini bersifat *open source* dan dibuat dengan menggunakan bahasa java dibawah lisensi GNU *Public License* dan *Rapid Miner* dapat dijalankan disistem operasi manapun. Dengan menggunakan *Rapid Miner*, tidak dibutuhkan kemampuan koding khusus, karena semua fasilitas sudah disediakan. *Rapid Miner* dikhususkan untuk penggunaan data mining.

METODE

Penelitian yang digunakan adalah model penelitian eksperimen. Bertujuan untuk melakukan prediksi Siswa SMA Bina Bangsa Mandiri dalam memperoleh Bantuan Dana Pendidikan dari sekolah, berdasarkan nilai akurasi dan evaluasi pada algoritma klasifikasi data mining. Penelitian ini menekankan pada teori-teori yang sudah ada, dilandasi oleh kerangka pemikiran pemecahan masalah seperti pada gambar 1.



Gambar 1. Kerangka Pemikiran Pemecahan Masalah

Penelitian ini dilakukan dengan menjalankan beberapa langkah proses penelitian yaitu :

1. Pengumpulan Data
2. Pengolahan Data Awal
3. Pengukuran Penelitian
4. Analisa hasil Penerapan Algoritma

HASIL DAN PEMBAHASAN

Faktor Exacta *11 (3):* 266-274, 2018
p-ISSN: 1979-276X
e- ISSN: 2502-339X
DOI : 10.30998/faktorexacta.v11i3.2777

Data yang digunakan bersumber dari data Siswa aktif SMA Bina Bangsa Mandiri dengan dengan jumlah kasus sebanyak 254 *record*, dan terdiri dari 37 atribut, seperti terlihat dalam gambar 2 dibawah ini:

Gambar 2. Potongan *Dataset*

Pengolahan Data Awal

Dalam pengujian ini menggunakan *rapid miner* dengan *operator 10-fold cross-validation* untuk mendapatkan hasil *accuracy* dan AUC pada setiap algoritma yang diuji menggunakan *dataset* siswa. Berikut adalah proses *Preprocessing.*

Gambar 3. Proses *Preprocessing*

Gambar 4. Potongan Hasil *Preprocessing*

Pengukuran Penelitian

Berdasarkan hasil olah data melalui aplikasi Rapid Miner, maka diperoleh *Confution Matrix* guna mengukur tingkat akurasinya, dari algoritma C4.5, yaitu sebesar **98,80%,** *Precision* sebesar **98,02%**, dan *Sensitivity* atau *Recall* sebesar **99,00%,** seperti yang dapat kita lihat pada gambar 5 dibawah ini.

Faktor Exacta *11 (3):* 266-274, 2018
p-ISSN: 1979-276X
e- ISSN: 2502-339X
DOI : 10.30998/faktorexacta.v11i3.2777

accuracy: 98.80% +/- 2.56% (mikro: 98.82%)

	true Layak	true Tidak	class precision
pred. Layak	99	2	98.02%
pred. Tidak	1	152	99.35%
class recall	99.00%	98.70%	

Gambar 5. Nilai Akurasi dari Algoritma C4.5

Nilai *accuracy* diperoleh dari rumus : $accuracy = \frac{TP+TN}{TP+FP+TN+FN}$

$$Accuracy = \frac{99+152}{99+1+152+2} = \frac{251}{254} = 0{,}9881$$

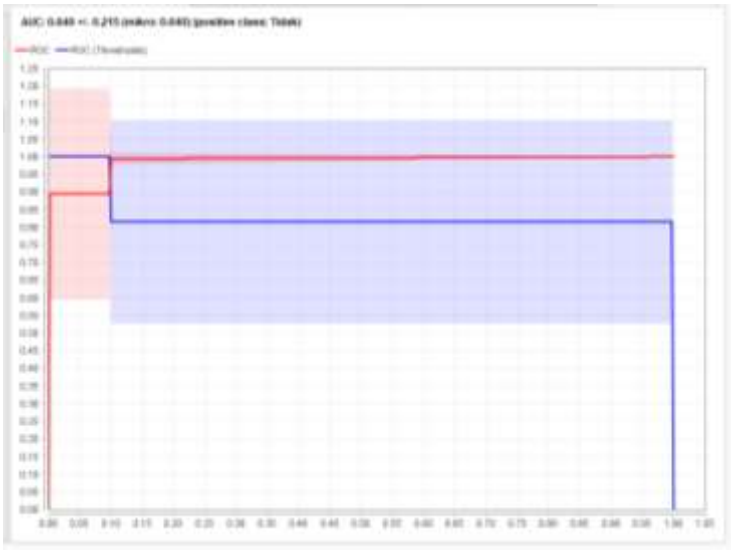
Mencari nilai *Sensitivity* atau *Recall* dengan rumus $sensitivity = \frac{TP}{TP+FN}$

$$Precision = \frac{99}{99+1} = \frac{99}{100} = 0{,}99$$

Mencari nilai *Precision* dengan rumus $precision = \frac{TP}{TP+FP}$

$$sensitivity = \frac{99}{99+2} = \frac{99}{101} = 0{,}9801$$

Area Under Curve (AUC) dari model Algoritma C4.5 dapat dilihat pada gambar 6 dibawah ini :



Gambar 6. *AUC* Algoritma C4.5

Penjelasan Area Under Curve (AUC) :

- Jika kurva yang dihasilkan mendekati garis baseline atau garis yang melintang dari titik 0,0, maka dapat dikaterikan jelek.

Faktor Exacta *11 (3):* 266-274, 2018
p-ISSN: 1979-276X
e- ISSN: 2502-339X
DOI : 10.30998/faktorexacta.v11i3.2777

- Sebaliknya jika kurva mendekati titik 0,1 atau diatasnya maka dikategorikan Bagus.

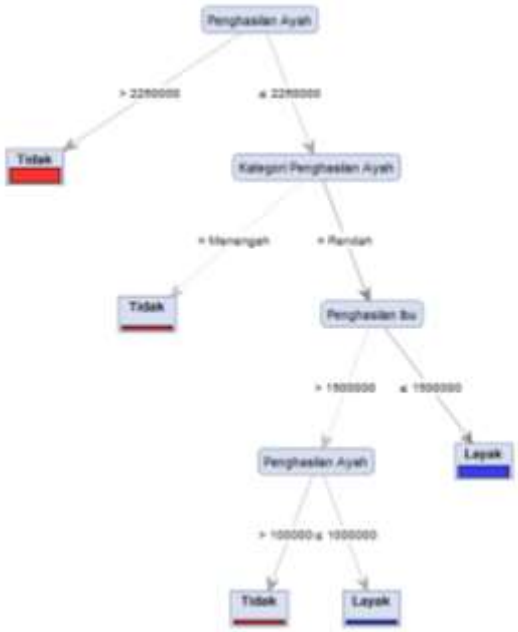
Tabel Evaluasi hasil pengolahan data dengan menggunakan metode yang diusulkan dapat dilihat pada tabel dibawah ini :

Tabel 2. Evaluasi Hasil Pengolahan data dengan Metode yang diusulkan

No	Metode	*Accuracy*	*Precision*	*Sensitivity* atau *Recall*
1	Algoritma C4.5	**98,80%**	**98,02%**	**99,00%**

Analisa Hasil Penerapan Algoritma

Melalui data yang disajikan pada table 1 di atas, telah menghasilkan sebuah model pohon keputusan, dapat dilihat pada gambar 7:



Gambar 7. Pohon Keputusan yang dihasilkan

Penjelasan dari Model Pohon Keputusan yang dihasilkan:

Yang menjadi atribut akar adalah Penghasilan ayah, dimana memiliki dua kelompok kelas atau *range* antara lain > 2250000 dengan <= 2250000. Diantara dua kelompok kelas tersebut, kelompok kelas penghasilan ayah >2250000 sudah mengklasifikasikan kasus kedalam keputusan Pilihan yang Tidak, sedangkan untuk kelompok kelas penghasilan ayah <= 2250000 itu perlu dilakukan perhitungan, guna menentukan node cabang selanjutnya. Begitu juga penjelasan untuk tiap node dibawahnya.

Berikut adalah *rule* yang dihasilkan dari pohon keputusan, dapat dilihat pada gambar 8 dibawah ini :

Tree

```
Penghasilan Ayah > 2250000: Tidak {Layak=0, Tidak=146}
Penghasilan Ayah ≤ 2250000
|   Kategori Penghasilan Ayah = Menengah: Tidak {Layak=0, Tidak=3}
|   Kategori Penghasilan Ayah = Rendah
|   |   Penghasilan Ibu > 1500000
|   |   |   Penghasilan Ayah > 1000000: Tidak {Layak=0, Tidak=5}
|   |   |   Penghasilan Ayah ≤ 1000000: Layak {Layak=4, Tidak=0}
|   |   Penghasilan Ibu ≤ 1500000: Layak {Layak=96, Tidak=0}
```

Gambar 8. *Rule* yang dihasilkan

PENUTUP

Simpulan

Dari hasil pengujian digunakan tes *Cros Validation* dan *Confusion Matrix* dan Kurva ROC. Algoritma C4.5 menghasilkan nilai *Accuracy* sebesar **98,80%**, nilai untuk *Precision* sebesar **98,02%**, dan nilai untuk *Sensitivity* atau *Recall* sebesar **99,00%**. Dengan demikian Algoritma C4.5 merupakan algoritma dan teknik terbaik untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan.

Saran

Saran untuk penelitian selanjutnya adalah dapat dilakukan sebuah komparasi dari beberapa Algoritma klasifikasi datamining, misal Algoritma C4.5 dengan Algoritma *Naïve Bayes* atau antara Algoritma C4.5, Algoritma *Naïve Bayes* dengan Algoritma *Random Fores*, sehingga dapat diketahui Algoritma yang terbaik dalam Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan.

DAFTAR PUSTAKA

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. *Chapman and Hall/CRC* (1st ed., Vol. 19). Chapman and Hall/CRC.

Bujlow, T., Riaz, T., & Pedersen, J. M. (2012). A method for classification of network traffic based on C5.0 machine learning algorithm. *2012 International Conference on Computing, Networking and Communications, ICNC'12*, 237–241. http://doi.org/10.1109/ICCNC.2012.6167418.

Dawson, C. W. (2009). *Projects in Computing and Information Systems A Student's Guide (2nd ed). Great Britain: Pearson Education*.

Gorunescu, Florin. (2011). *Data Mining: Concepts and Techniques*. Verlag berlin Heidelberg: Springer.

Han, J., & Kamber, M. (2007). *Data Mining Concepts and Techniques.* San Fransisco: Mofgan Kaufan Publisher.

Han Jiawei,Kamber, M, "*Data Mining: Concepts and Techniques*," New York: Morgan Kaufmann Publishers, *2001*.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*.

Larose, D. T. (2005). *Discovering Knowledge in Data*. New Jersey: John Willey & Sons, Inc.

Larose Daniel T (2006). *Data Mining Methodes* and *Models*. Hoboken. Wiley-Interscience. : John Willey & Sons, Inc.

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications - A decade review from 2000 to 2011. *Expert Systems with Applications*, *39*(12), 11303–11311. http://doi.org/10.1016/j.eswa.2012.02.063.

Turban, E., dkk, 2005, Decicion Support Systems and Intelligent Systems, Andi Offset.