Penerapan Fitur Seleksi Forward Selection Menggunakan Algoritma Naive Bayes Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Universitas AKI Semarang

Bondhan Arya Purnanditya[1], Ahmad Zainul Fanani[2]
[1,2]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Jl. Nakula 1 No. 5-11 Semarang 50131, Telp. (024) 3520165 Fax : 3569684
E-mail : bondhan.aryafmd@yahoo.com[1], a.zainul.fanani@dsn.dinus.ac.id[2]

ABSTRAK

Universitas adalah suatu institusi pendidikan tinggi dan penelitian yang memberikan gelar akademik dalam berbagai bidang. Universitas didirikan untuk mengarahkan lulusannya menjadi tenaga profesional, siap kerja, tenaga pendidikan, atau bahkan peneliti. Dataset status kelulusan mahasiswa dari IAsol Dataset adalah dataset yang diambil dari IAsol khususnya pada fakultas ilmu komputer. Algoritma Naive Bayes diketahui bisa memecahkan masalah data dataset dengan dimensi data yang besar dan bersifat Class Imbalance dengan hasil akurasi hanya 95.83%. Metode Forward Selection digunakan untuk mereduksi dimensi dataset yang besar dan dapat membantu meningkatkan hasil akurasi klasifikasi Naïve Bayes. Algoritma Naïve Bayes dengan Forward Selection sebagai fitur seleksi terbukti akurat dan efektif dalam mengklasifikasikan kelulusan mahasiswa dengan hasil akurasi 99.17% dan termasuk dalam kategori "Kappa excellent".

Kata kunci: *Data Mining, Klasifikasi, Naive Bayes, Forward Selection, Class*

ABSTRACT

Imbalance.University is an institution of higher education and research that provide academic degrees in various fields. University was established to direct graduates become professionals, ready to work, education personnel, or even researchers. Dataset graduation status of students of IAsol dataset is a dataset taken from IAsol especially in computer science faculty. Naive Bayes algorithm is known to solve the problem of data-dimensional dataset with large data and are Class Imbalance with the results of only 95.83% accuracy. Forward Selection method used to reduce the dimensions of large datasets and can help improve the accuracy of results Naïve Bayes classification. Naïve Bayes algorithm with Forward Selection as feature selection proved accurate and effective in classifying graduation with 99.17% accuracy and results are included in the category of "Kappa excellent".

Keywords: *Data Mining, Classification, Naive Bayes, Forward Selection, Class Imbalance.*

1. PENDAHULUAN

Universitas dalam pendidikan di Indonesia merupakan salah satu bentuk perguruan tinggi selain akademi, institut, politeknik, dan sekolah tinggi [1]. Universitas terdiri atas sejumlah fakultas yang menyelenggarakan pendidikan akademik dan pendidikan vokasi pada sejumlah ilmu pengetahuan, teknologi, seni dan jika memenuhi syarat dapat menyelenggarakan pendidikan profesi. Universitas adalah suatu institusi pendidikan tinggi dan penelitian, yang memberikan gelar akademik dalam berbagai bidang. Universitas didirikan untuk mengarahkan lulusannya menjadi

tenaga profesional, siap kerja, tenaga pendidikan, atau bahkan peneliti. Pada umumnya program yang ditawarkan di salah satu Universitas adalah program pendidikan sarjana dan pascasarjana [2]. Didalam suatu universitas terdapat beberapa fakultas-fakultas diantaranya fakultas ilmu komputer, fakultas ekonomi, fakultas bahasa dan fakultas lainnya.

Berdasarkan berlimpahnya data mahasiswa dan data jumlah kelulusan mahasiswa, informasi yang tersembunyi dapat diketahui dengan cara melakukan pengolahan terhadap data mahasiswa sehingga berguna bagi pihak universitas [5]. Pengolahan data mahasiswa perlu dilakukan untuk mengetahui informasi penting berupa pengetahuan baru (knowledge Discovery), misalnya informasi mengenai pengklasifikasian data mahasiswa berdasarkan profil dan data akademik. Pengetahuan baru tersebut dapat membantu pihak universitas untuk melakukan klasifikasi mengenai tingkat kelulusan mahasiswa guna menetukan strategi untuk meningkatkan kelulusan pada tahun-tahun berikutnya.

Penelitian ini akan melakukan pengklasifikasian berdasarkan dataset IAsol yang didapat dari Universitas Abadi Karya Indonesia khususnya di Fakultas Ilmu Komputer pada tahun ajaran 2008 sampai 2011. Atribut yang akan digunakan dalam melakukan klasifikasi kelulusan adalah Nomor Induk Mahasiswa (NIM), nama, jurusan, umur, jenis kelamin, daerah asal, status pernikahan, status pekerjaan, kelompok atau jenis beasiswa, indeks prestasi dari semester 1 sampai dengan semester 9, IPK, jumlah sks yang ditempuh dan jenis konsentrasi jalur peminatan.

Berbagai algoritma klasifikasi Data Mining telah banyak diterapkan untuk membantu mengklasifikasikan penentuan status kelulusan salah satunya menggunakan Naïve bayes. Naïve bayes diketahui memiliki kecepatan komputasi yang sangat tinggi, mampu menangani masalah data dataset yang berdimensi besar dan dataset yang bersifat Class Imbalance [8] [9] [10] [11] [12]. Pada penelitian kali ini selain mendapatkan nilai akurasi yang baik juga bertujuan mendapatkan model atribut yang berpengaruh dengan cara menerapkan Feature Selection.
Feature Selection adalah salah satu cara untuk menentukan atribut yang paling berpengaruh di dalam dataset. Feature Selection berperan memilih subset yang tepat dari set fitur asli, karena tidak semua fitur/atribut relevan dengan masalah [13]. Bahkan beberapa dari fitur atau atribut tersebut mengganggu dan dapat mengurangi akurasi. Noisy Features atau fitur yang tidak terpakai tersebut harus dihapus untuk meningkatkan akurasi. Selain itu dengan fitur atau atribut yang banyak akan memperlambat proses komputasi.
Wraper Feature Selection terdiri dari Forward Selection, Backward Elimination dan Stepwise Selection. Forward Selection dan Stepwise Selection memiliki hasil yang lebih memuaskan dibandingkan dengan proses Backward Elimination. Forward Selection juga memerlukan waktu komputasi yang relatif lebih pendek dibandingkan dengan Backward Elimination maupun dengan Stepwise Selection.

Pada penelitian ini akan menggunakan Forward Selection. Forward Selection atau seleksi kedepan dalam analisisnya pemilihan ke depan di mulai dengan tidak ada prediktor dalam model untuk membantu meningkatkan hasil akurasi dan menentukan atribut yang berpengaruh.

2. METODE

Jenis penelitian yang dilaksanakan ini merupakan penelitian eksperimen. Selain itu data yang digunakan adalah data kualitatif. Data kualitatif adalah data yang berupa kalimat. Data-data dalam penelitian berasal dari UNAKI.

2.1 Pengumpulan Data

Pengumpulan data pada penelitian ini meliputi: studi literatur yang digunakan sebagai referensi dalam penelitian bisa berupa buku, jurnal dan karya ilmiah yang relevan dengan algoritma klasifikasi data mining. Tahap ini dilakukan sebagai langkah awal dari suatu penelitian. Untuk memperoleh data yang benar-benar akurat, maka penentuan jenis dan sumber data sangatlah penting. Sumber data pada penelitian ini adalah dataset yang didapat dari IAsol UNAKI [7] khususnya di Fakultas Ilmu Komputer pada tahun ajaran 2008 sampai 2011. Atribut yang akan digunakan dalam melakukan klasifikasi kelulusan adalah Nomor Induk Mahasiswa (NIM), nama, jurusan, umur, jenis kelamin, daerah asal, status pernikahan, status pekerjaan, kelompok atau jenis beasiswa, indeks prestasi dari semester 1 sampai dengan semester 9, IPK, jumlah sks yang ditempuh dan jenis konsentrasi jalur peminatan.

2.2 Teknik Analisis Data

Tahap pengolahan awal data dilakukan untuk mempersiapkan data yang benar-benar valid sebelum diproses pada tahap berikutnya namun tidak semua data dapat digunakan dan tidak semua atribut digunakan karena harus melalui beberapa tahap pengolahan awal data (preparation data). Jumlah data awal yang diperoleh dari pengumpulan 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, menurut Vercellis [26] dilakukan beberapa teknik:

1. Data integration and transformation, untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal. Data ditransformasikan kedalam software RapidMiner.

2. Data size reduction, untuk memperoleh data set dengan jumlah atribut dan record yang lebih sedikit tetapi tetap bersifat informatif.

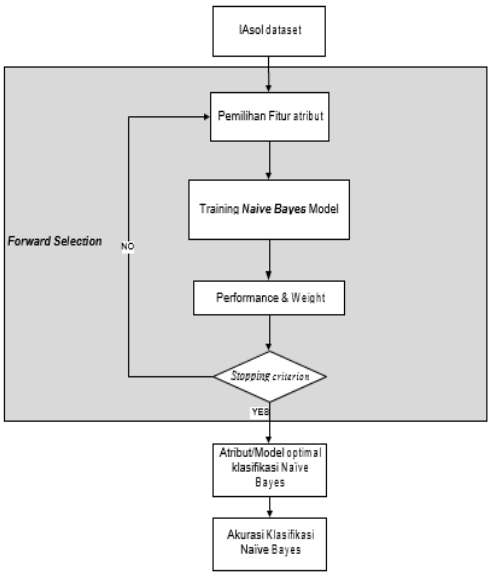
Pada penelitian kali ini tahapan yang dilakukan hanya transformasi data yaitu merubah beberapa tipe atribut data agar dikenali oleh RapidMiner.

NIM	Integer
Nama	Polynominal
Jurusan	Binominal
Umur	Integer
Jenis Kelamin	Binominal
Daerah Asal	Polynominal
Status Pernikahan	Binominal
Status Pekerjaan	Binominal
Kelompok	Polynominal
IP Semester 1	Numeric
IP Semester 2	Numeric
IP Semester 3	Numeric
IP Semester 4	Numeric
IP Semester 5	Numeric
IP Semester 6	Numeric
IP Semester 7	Numeric
IP Semester 8	Numeric
IP Semester 9	Numeric
IPK	Numeric
SKS	Integer

Konsentrasi	Polynominal
Status	Polynominal

Tabel 2.1 Tipe atribut data

2.3 Metode Penelitian



Gambar 2.1 Tahapan proposed method

Tahap ini akan membahas metode yang akan digunakan untuk penelitian. Berikut ini adalah tahap yang akan dilakukan dalam penelitian. Tahapan dilakukan mengikuti langkah-langkah metode Forward Selection dengan algoritma Naïve Bayes yaitu:

Dataset dari Iasol UNAKI diseleksi fitur menggunakan Forward Selection, Metode Forward Selection adalah pemodelan dimulai dari nol peubah (empty model).
Pemilihan fitur seleksi forward selection diuji menggunakan training atau metode Naive Bayes.
Dari training Naive Bayes yang diujikan mendapatkan hasil dan pembobotan.

Apabila proses tersebut lolos maka akan mendapatkan suatu atribut/model yang optimal dari klasifikasi Naive Bayes.

Sedangkan bila proses tersebut berhenti pada stopping criterion maka proses tersebut diulang dari awal (pemilihan fitur seleksi forward selection) sampai mendapatkan atribut/model optimal.
Setelah mendapatkan atribut/model yang optimal pada klasifikasi Naive Bayes maka akan muncul hasil akurasi dari klasifikasi Naive Bayes yang sudah di fitur seleksi.
Tahap ini akan membahas metode yang akan digunakan untuk penelitian nanti. Berikut ini adalah tahap yang akan dilakukan dalam penelitian. Seleksi fitur digunakan sebagai input untuk proses klasifikasi. Seleksi fitur dilakukan dengan mengambil sebagian variabel pada seluruh atribut yang ada pada data untuk dijadikan atribut penentu dalam melakukan pemberian keputusan. Dataset diseleksi fitur menggunakan Forward Selection, proses selanjutnya adalah melakukan klasifikasi menggunakan algoritma Naïve Bayes, hasil proses klasifikasi di evaluasi dengan menggunakan Confussion Matrix dan Kappa untuk mengukur performan atau tingkat akurasi.

2.4 Pengujian Model/Metode

Pada tahapan ini menjelaskan tentang teknik pengujian yang digunakan. Tahap modeling untuk mengklasifikasikan status kelulusan dengan menggunakan dua metode yaitu algoritma Naïve Bayes dan Forward Selection-Naïve Bayes. Proses eksperimen dan pengujian model menggunakan dataset IAsol [7]. Metode eksperimen dan pengujian ini mengikuti cara pengklasifikasian menggunakan RapidMiner.

2.5 Evaluasi Dan Validasi Hasil

Pada tahap ini akan dibahas tentang hasil evaluasi dari eksperimen yang telah dilakukan. Model yang terbentuk akan diuji dengan menggunakan

Confusion Matrix untuk mengetahui tingkat akurasi. Confusion Matrix akan menggambarkan hasil akurasi mulai dari prediksi positif yang benar, predisksi positif yang salah, prediksi negative yang benar, dan prediksi negative yang salah. Akurasi akan dihitung dari seluruh prediksi yang benar (baik prediksi positif dan negatif). Semakin tinggi nilai akurasi, semakin baik pula model yang dihasilkan.

Pengujian juga diukur dengan menggunakan Kappa, semakin tinggi nilai Kappa, maka semakin baik pula model klasifikasi yang terbentuk

3. HASIL DAN PEMBAHASAN

Pada penelitian ini menguji keakuratan klasifikasi kelulusan mahasiswa dengan menggunakan algoritma Naïve Bayes, setelah itu Naïve Bayes dengan Forward Selection sebagai fitur seleksi. Penelitian ini menggunakan dataset yang diambil dari IAsol Dataset yaitu dataset kelulusan mahasiswa yang memiliki 3 class atau 3 kategori kelulusan, dengan data yang besar (memiliki 240 record dan 21 attribute) serta bersifat class imbalance.

3.1 Algoritma Naïve Bayes

Naïve Bayes adalah metode yang baik karena mudah dibuat, tidak membutuhkan skema estimasi parameter perulangan yang rumit, ini berarti bisa diaplikasikan untuk dataset berukuran besar [19].

Berikut teorema bayes :

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

Berikut rumus Naive Bayes :

$$P(X|H) = P(H|X)P(H)$$

Keterangan :
X: Data dengan class yang belum diketahui
H: Hipotesis data *x* merupakan suatu class spesifik
P(H|X): Probabilitas hipotesis *H* berdasarkan 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*

3.2 Evaluasi Naïve Bayes dengan data sampel

Pengujian menggunakan data sampel yang diambil dari IAsol dataset dengan: 2 label class (tepat dan terlambat), 10 record (7 class tepat dan 3 class terlambat) dan 21 attribute seperti yang dapat dilihat pada halaman lampiran.
Berikut ini adalah contoh perhitungan mencari nilai akurasi dari atribut kelompok dengan menggunakan metode Cross-Validation (X-Validation).
Training 1:

Status Kelulusan	Kelompok
Tepat	Akademik
Tepat	Reguler
Tepat	Reguler
Tepat	Akademik
Tepat	Reguler
Tepat	Akademik
Tepat	Reguler
Terlambat	Parsial
Terlambat	Parsial

Tabel 3.1 data training cross validation naïve bayes

Dari data diatas didapatkan Probabilitas kelas:
P(Tepat) = 7/9 = 0.777777777
P(Terlambat) = 2/9 = 0.222222222
Dari data diatas didapatkan Probabilitas kelompok terhadap masing masing kelas:
P(Akademik|Tepat) = 3/7 = 0.428571429

P(Reguler|Tepat) = 4/7 = 0.571428572
P(Parsial|Tepat) = 0/7 = 0
P(Akademik|Terlambat) = 0/2 = 0
P(Reguler|Terlambat) = 0/2 = 0
P(Parsial|Terlambat) = 2/2 =1

Testing 1:
Data testing dari status kelulusan dengan kelompok parsial:
Prediction parsial: P(X|Tepat) = 0/7 = 0
P(X|Terlambat) = 2/2 = 1
Perhitungan dilakukan 10 kali sampai *training 10* dan *testing 10* sesuai metode *Cross-Validation* (*X-Validation).*
Dari hasil klasifikasi menggunakan data sample *(2 label class. 10 record dan 21 attribute)* dengan metode *Naïve Bayes* diperoleh hasil nilai akurasi sebesar 70.00%, berikut ini hasil perhitungannya seperti dapat dilihat pada gambar 4.1.

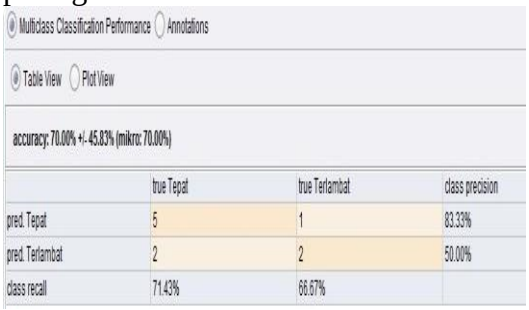

Multiclass Classification Performance Annotations

Table View Plot View

accuracy: 70.00% +/- 45.83% (mikro: 70.00%)

	true Tepat	true Terlambat	class precision
pred. Tepat	5	1	83.33%
pred. Terlambat	2	2	50.00%
class recall	71.43%	66.67%	

Gambar 3.1 Validasi Naïve bayes data Sampel

$$= \frac{5+2}{5+1+2+2}$$

= 0.7

= 70%

4. KESIMPULAN DAN SARAN

Algoritma Naive Bayes terbukti efektif dalam mengklasifikasikan status kelulusan mahasiswa dari dataset dengan dimensi data yang besar dan memiliki keadaan kelas yang tidak seimbang antara kelas yang satu dengan kelas yang lain atau bersifat class imbalance.

Metode Forward Selection dapat mereduksi dimensi dataset yang besar dan dapat membantu meningkatkan hasil akurasi klasifikasi Naïve Bayes.
Dalam hal ini Naive Bayes memanfaatkan fungsi seleksi fitur dari Forward Selection untuk pemilihan atribut data dengan karakteristik data itu sendiri, dan meningkatkan ketepatan klasifikasi Naïve Bayes.
Forward Selection berbasis Naive Bayes lebih akurat dan efektif dalam mengklasifikasikan status kelulusan mahasiswa dari dataset yang bersifat class imbalance dengan data yang besar dengan hasil akurasi 99.17% dan termasuk dalam kategori "Kappa excellent". Dengan memperoleh atribut yang berpengaruh yaitu: Kelompok, IP Semester 1, Semester 3, IP Semester 9. Dibanding dengan menggunakan algoritma Naive Bayes saja dengan hasil akurasi 95.83%.

4.1 Saran

Metode Forward Selection berbasis Naive Bayes terbukti akurat dalam klasifikasi status kelulusan mahasiswa dari dataset yang bersifat class imbalance dengan dimensi data yang besar, tetapi dalam penelitian ini terdapat beberapa saran dalam pengembangannya antara lain prosedur ini tidak selalu mengarahkan ke model pemilihan atribut yang terbaik. Forward Selection berbasis Naive Bayes hanya mempertimbangkan sebuah subset kecil dari semua model-model yang mungkin, sehingga resiko melewatkan atau kehilangan model terbaik akan bertambah, seiring dengan penambahan jumlah variabel bebas.
Membantu administrasi perguruan tinggi untuk memberikan peringatan dini dan pembimbingan awal bagi mahasiswa yang kemungkinan tidak lulus tepat waktu dan membantu

perguruan tinggi dalam membuat kebijakan untuk bisa meningkatkan kelulusan mahasiswa

Penelitian ini dapat dikembangkan dengan metode klasifikasi data mining lainnya, penggunaan metode fitur seleksi atau metode optimasi lainnya yang dapat mengatasi masalah dimensi data yang besar, class imbalance dan multiclass seperti pada penelitian ini.

DAFTAR PUSTAKA

[1] Undang-Undang Republik Indonesia Nomor 2 Tahun 1989 tentang Sistem Pendidikan Nasional.

[2] Romi Satria Wahono, Dapat Apa Sih Dari Universitas? Bandung: Zip Book, 2009.

[3] Jennifer Streubel , "What Is Computer Science," Department of Computer Science, Boston, 2003.

[4] David M Kroenke, Experiencing MIS.: Prentice Hall, Upper Saddle River, NJ, 2008.

[5] Johan Oscar Ong, "Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University," Jurnal Ilmiah Teknik Industri, Juni 2013.

[6] http://www.unaki.ac.id/.

[7] http://iasol.unaki.ac.id:9090/IasolWeb/Login.aspx?ReturnUrl=%2fIasolWeb%2fdefault.aspx.

[8] Yi Liu , Lei Wei , and Peng Wang, "Regional Style Automatic Identification for Chinese folk Songs," 2009.

[9] Christopher DeCoro, Zafer Barutcuoglu, and Rebecca Fiebrink, "Bayesian Aggregation For Hierarchical Genre Classification," in Austrian Computer Society (OCG), 2007.

[10] Alfa Saleh , "Penerapan Data Mining Dengan Metode Klasifikasi Naive Bayes Untuk Memprediksi Kelulusan Mahasiswa Dalam Mengikuti Test English Proficiency Test".

[11] Mujib Ridwan , Hadi Suyono , and M. Sarosa , "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," Juni 2013.

[12] Mohamad Fajarianditya Nugroho, Romi Satria Wahono, and Vincent Suhartono , "Penerapan Metode Forward Selection untuk Fitur Seleksi Pada Klasifikasi Genre Musik Menggunakan Algoritma Naive Bayes," Udinus, Mkom Thesis 2013.

[13] Oded Maimon and Lior Rokach, Data Mining and Knowledge Discovery Handbook , 2nd ed.: Springer, 2010.

[14] Carlos N. Silla Jr, Alessandro L. Koerich, and Celso A. A. Kaestner, "Feature Selection in Automatic Music Genre Classification," in Tenth IEEE International Symposium on Multimedia, 2008.

[15] L. Ladha and T. Deepa, "Feature Selection Methods And Algorithms," International Journal on Computer Science and Engineering (IJCSE), vol. 3, p. 5, May 2011.

[16] Mark A. Hall and Geoffrey Holmes , "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining," IEEE Transactions On Knowledge And Data Engineering, vol. 15, p. 3, May/June 2003.

[17] Ian H Witten, Eibe Frank, and Mark A Hall, Data Mining Practical Machine Learning Tools and Techniques, 3rd ed.: Morgan Kaufmann , 2011.

[18] Florin Gorunescu, Data Mining: Concepts, Model and Techniques, Prof. Janusz Kacprzyk and Prof. Lakhmi C. Jain, Eds. Berlin,

Jerman: Springer, 2011, vol. 12.
[19] Xindong Wu and Vipin Kumar, The top ten Algorithms in Data Mining.: Taylor & Francis Group, LLC, 2009.
[20] Budi Santoso, Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis, 1st ed. Yogyakarta, Indonesia, 2007.
[21] Jiawei Han, Data Mining Concept And Technique, 2nd ed., Asma Stephan, Ed. Champaign, United States of America: Multiscience Press, 2007.
[22] Jacob Cohen, "A Coefficient Of Agreement For Nominal Scale ," 1960.
[23] Joseph L. Fleiss , "Measuring Nominal Scale Agreement Among Many Raters," 1971.
[24] Mikael Berndtsson, Jörgen Hansson, Björn Olsson, and Björn Lundell, Thesis Projects A Guide for Students in Computer Science and Information Systems, 2nd ed. London: Springer, 2008.
[25] Christian W Dawson, Projects in Computing and Information Systems A Student's Guide, 2nd ed. England: Pearson Education, 2009.
[26] Carlo Vercellis, Business Intelligent: Data Mining and Optimization for Decision Making. Southern Gate, Chichester: John Willey & Sons, Ltd., 2009.