JUSTIN
Jurnal Sistem dan Teknologi Informasi
p-ISSN : 2460-3562 / e-ISSN : 2620-8989
Vol. 07, No. 03, Juli 2019

Prediksi Ketepatan Kelulusan Mahasiswa Diploma dengan Komparasi Algoritma Klasifikasi

Muhammad Sony Maulana[#1], Raja Sabarudin[#2], Wahyu Nugraha[#3]

[#]*Program Studi Sistem Informasi Kampus Kota Pontianak, Universitas Bina Sarana Informatika*
Jalan Abdurahman Saleh No.18 A Pontianak

[1]muhammad.sony.mom@bsi.ac.id
[2]raja.rjd@bsi.ac.id
[3]wahyu.whn@bsi.ac.id

***Abstrak*— AMIK BSI Pontianak merupakan salah satu perguruan tinggi swasta yang memiliki jumlah mahasiswa yang banyak, namun dalam perjalanannya masih terdapat permasalahan yang setiap tahun nya terjadi yaitu permasalahan jumlah kelulusan mahasiswa yang tepat waktu dan terlambat. Jumlah mahasiswa yang lulus tepat waktu menjadi indikator efektifitas dari sebuah perguruan tinggi baik negeri dan swasta. Perguruan tinggi perlu mendeteksi perilaku dari mahasiswa aktif sehingga dapat dilihat faktor yang menyebabkan mahasiswa tidak lulus tepat waktu. Pada penelitian ini, akan mengkomparasikan atau membandingkan 5 metode *data mining* untuk menentukan metode mana yang paling optimal dalam menentukan ketepatan kelulusan mahasiswa dengan teknik pengujian *T-Test*, metode yang dibandingkan adalah metode *Decision Tree*, *Naive Bayes*, K-NN, *Rule Induction*, dan *Random Forest*. Hasil dari penelitian ini menghasilkan bahwa algoritma *Rule Induction* dan C4.5 adalah metode yang paling optimal performanya dalam menentukan ketepatan kelulusan mahasiswa diploma AMIK BSI Pontianak.**

***Kata kunci*— Komparasi, Algoritma Klasifikasi, Ketepatan Kelulusan Mahasiswa, data mining, pengujian T-Test**

I. PENDAHULUAN

Berkembangnya penggunaan teknologi informasi dan komputer dalam bidang pengelolaan data menyebabkan akumulasi data dalam jumlah sangat besar di berbagai macam perusahaan. Apalagi dengan semakin berkembangnya pengetahuan mengenai data yang besar tersebut menjadi informasi yang lebih berguna bagi pengelola data untuk lebih dioptimalkan dalam menunjang keputusan-keputussan bisnis yang menguntungan. Data dalam jumlah besar kadangkala belum dioptimalkan untuk memperoleh informasi yang lebih mendalam untuk mendukung tujuan-tujuan strategis yang diperlukan sehingga seolah-olah dibiarkan begitu saja.

Demikian halnya data dalam institusi perguruan tinggi seperti AMIK BSI Pontianak yang menyimpan kumpulan data yang banyak. Sejak berdiri hingga sekarang terdapat kurang lebih 5000 mahasiswa sehingga terdapat banyak data yang bisa digali. Peranan sebagai institusi pendidikan diharapkan menyelenggarakan pendidikan yang berkualitas bagi mahasiswa sehingga menghasilkan sumber daya manusia yang berilmu, kreatif dan berkualitas. Perlu digali informasi yang bisa digunakan untuk lebih meningkatkan daya saing yang salah satunya untuk meningkatkan prestasi akademik mahasiswa. Prestasi akademik sebagai keberhasilan seorang mahasiswa tidak terlepas dari latar belakang mahasiswa itu sendiri disamping sistem dan iklim belajar mengajar yang tercipta di lingkungan pendidikannya. IPK yang baik tentunya membuat target masa studi tercapai dengan kualitas yang bagus. Masa studi yang tepat waktu mendorong berkurangnya penumpukan mahasiswa di semester akhir yang bisa mengakibatkan ratio dan kualitas yang tidak baik. Prestasi Akademik biasanya diukur melalui Indek Prestasi Akademik (IPK). Keberhasilan dalam memperoleh IPK yang tinggi biasanya dipengaruhi oleh banyak faktor seperti jenis kelamin, status mahasiswa, umur, status nikah, nilai indeks prestasi semester(IPS), dan lain-lain.

Dalam mengolah data mahasiswa untuk prediksi kelulusan telah diselesaikan oleh Karamouiz dan Vrettos dengan menggunakan metode neural network [9], Qudri dan Kalyankar dengan metode decision tree [6], Suhartini dan Ernastuti dengan metode C4.5 dan naïve bayes, [2], Hastuti dengan komparasi metode Logistic Regression, Decision Tree, Naïve Bayes, Neural Network [5] dan Tahyudin, Utami dan Amborowati dengan mengkomparasi algoritma decision tree, *naïve bayes*, ANN, Support Vector Machine (SVM) dan Logistic Regression (LR) [7].

Pada penelitian ini yang akan dilakukan adalah membandingkan 5 buah metode algoritma data mining untuk menentukan metode mana yang paling optimal dalam menentukan ketepatan kelulusan mahasiswa dengan bantuan teknik pengujian T-Test dan dengan menggunakan sofware

Submitted 20-06-2019; Revised 16-07-2019; Accepted 29-07-2019

Rapid Miner dan dengan menggunakan dataset mahasiswa AMIK BSI Pontianak angkatan 2013/2017 Prodi Diploma Manajemen Informatika sebanyak 349 *record*.

II. PENELITIAN TERKAIT

Penelitian yang dilakukan oleh Karamouiz dan Vretoz pada tahun 2009 dengan judul Sentivity Analysis of Neural Network for Identifying the Factors for Collage Students Success. Masalah yang yang dikaji adalah tingkat kelulusan dianggap sebagai indikator efektivitas suatu lembaga institusi, Metode yang digunakan adalah NN (Neural Network). Dari hasil data training yang dilakukan diperoleh kategori yang lulus adalah 86.04% dan training data untuk kategori yang tidak sukses adalah 68.21%, dan error yang diperoleh untuk kedua kategori sebut adalah 0.18%.

Penelitian yang dilakukan oleh Qudri dan Kalanyar pada tahun 2010 dengan judul Drop Out Feature of Student Data for Academic Performance Using Decision Tree techniques. Masalah dalam penelitiaanya adalah prestasi akademik siswa sangat penting bagi lembaga pendidikan karena program-program strategis dapat direncanakan untuk meningkatkan atau mempertahankan prestasi siswa selama periode mereka studi di lembaga. Metode yang digunakan adalah Decision Tree, yakni algoritma J4.8. Hasil penelitian ini adalah sebuah pohon keputusan yang dapat dijadikan rule bagi prediksi siswa yang putus sekolah.

Penelitian yang dilakukan oleh Suhartina dan Ernastuti pada tahun 2010 dengan judul dengan judul Graduation Prediction of Gunadarma University Students Using Algorithm and Naive Bayes C4.5 Algorithm. Permasalahannya adalah banyaknya mahasiswa yang tidak lulus tepat waktu. Untuk mengetahui tingkat kelulusan mahasiswa dalam satu tahun ajaran dapat dilakukan suatu prediksi berdasarkan data-data mahasiswa pada tahun ajaran pertama. Algoritma yang digunakan adalah C45 dan naïve bayes. Hasil dari penelitian ini adalah akurasi untuk metode naïve bayes adalah 80,85% dengan presentasi kesalahan 19,05% Akurasi ketepatan hasil prediksi C4.5 85.7%, dan presentasi kesalahannya adalah 14,3%.

Penelitian yang dilakukan oleh Hastuti pada tahun 2012 dengan judul analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif. Permasalahannya adalah mahasiswa non aktif adalah mahasiswa yang berhenti studi dan tidak melakukan registrasi administratif. Mahasiswa yang memiliki status non aktif memiliki kecenderungan untuk drop out. Metode yang digunakan adalah Logistic Regression, Decision Tree, Naïve Bayes, Neural Network. Hasilnya adalah akurasi Logistic Regression 81,64%, Decision Tree 95,29%, Naïve Bayes 93,47%, dan Neural Network 94,59%.

Penelitian yang dilakukan oleh Tahyudin, Utami dan Amborowati pada tahun 2013 dengan judul Comparing Clasification Algorithm Of Data Mining to Predict the Graduation Students on Time. Permasalahannya adalah persentase mahasiswa yang lulus tepat waktu adalah salah satu unsur yang mempengaruhi akreditas program studi. Metode yang digunakan adalah mengkomparasi algoritma decision tree, naïve bayes, ANN, Support Vector Machine (SVM) dan Logistic Regression (LR). Hasilnya adalah akurasi algoritma decision tree 80,01%, *naïve bayes* 75,16%, ANN 100%, SVM 100%, dan LR 100%.

Pada penelitian ini akan dilakukan komparasi 5 buah metode algoritma data mining yaitu metode Decision Tree, *Naive Bayes*, K-NN, Rule Induction, dan Random Forest untuk menentukan metode mana yang paling optimal dalam menentukan ketepatan kelulusan mahasiswa dengan bantuan teknik pengujian komparasi T-Test serta dengan menggunakan software Rapid Miner.

III. METODE TERKAIT

Penelitian ini menggunakan metode penelitian komparatif. Metode analisis yang akan dipakai adalah metode *data mining* klasifikasi *dengan tools software rapid miner studio*. Proses penelitian yang akan dilakukan dapat dilihat pada gambar 1.

A. Dataset

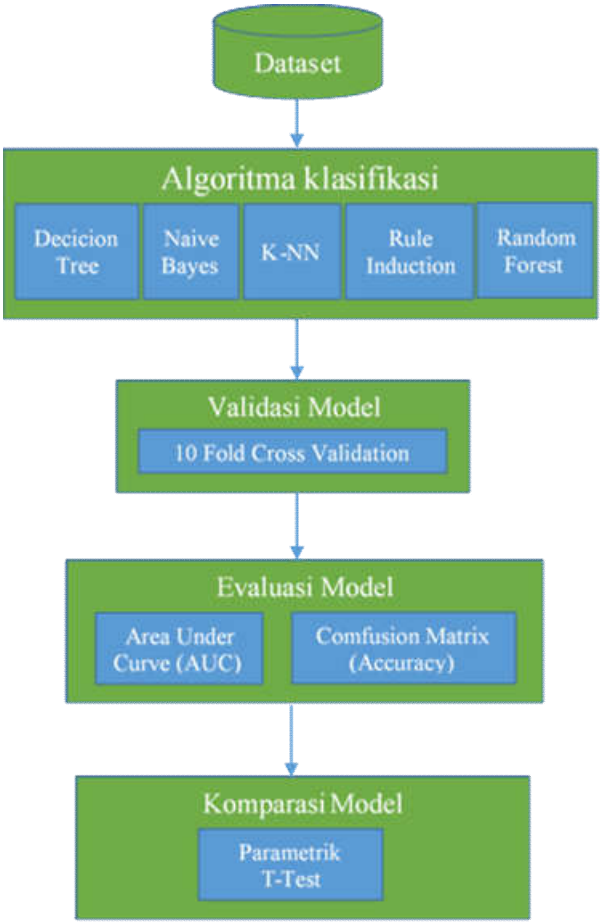
Dataset yang digunakan dalam penelitian ini adalah dataset AMIK BSI Pontianak angkatan 2013/2017 Prodi D3 Manajemen Informatika sebanyak 349 *Record*. Data Atribut yang digunakan terdapat pada Tabel 1.

TABEL I
TABEL DATA ATRIBUT

Atribut	**Keterangan**
Nama	Nama Mahasiswa
Jenis Kelamin	Laki-laki, Perempuan
Status Mahasiswa	Bekerja, Mahasiswa
Umur	Umur Mahasiswa
Status Nikah	Menikah, Belum Menikah
IPS1	Nilai Semester
IPS2	Nilai Semester
IPS3	Nilai Semester
IPS4	Nilai Semester
IPS5	Nilai Semester
IPS6	Nilai Semester
IPK	Nilai Komulatif Semester
Status Kelulusan	Tepat, Terlambat

B. Rancangan Algoritma Klasifikasi

Algoritma Klasifikasi ini digunakan untuk membandingkan performansi dari model-model klasifikasi untuk memprediksi ketepatan dalam kelulusan mahasiswa. Dalam penelitian ini telah dipilih 5 model klasifikasi yaitu *Decision Tree, Naive Bayes, K-NN, Rule Induction*, dan *Random Forest.*



Gambar 1. Alur Penelitian Komparasi dengan Model Klasifikasi untuk Prediksi Ketepatan Kelulusan Mahasiswa.

C. *Validasi Model*

Validasi model dalam penelitian ini menggunakan *Cross Validation* adalah teknik validasi dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi [4]. Dengan menggunakan *cross validation* akan dilakukan percobaan sebanyak k. Data yang digunakan dalam percobaan ini adalah data *training* untuk mencari nilai *error rate* secara keseluruhan. Secara umum pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi. Dalam penelitian ini nilai k yang digunakan berjumlah 10 atau 10-*fold Cross Validations*. Hasil dari berbagai percobaan yang ekstensif dan pembuktian teoritis, menunjukkan bahwa 10-*fold cross-validation* adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat. Karena 10-*fold cross-validation* akan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian, seperti terlihat pada tabel 2. Penelitian ini menggunakan *k-fold cross validation* karena metode ini telah menjadi metode validasi standar dan *state-of-the-art*, bahkan beberapa tes juga menunjukkan bahwa penggunaan stratified k-fold cross-validation sedikit meningkatkan hasil dari pengujian model [10].

TABEL III
TABEL STRATIFIED 10 FOLD CROSS VALIDATION

n-Validasi	Pembagian *Dataset*									
1	■	■								
2		■	■							
3			■	■						
4				■	■					
5					■	■				
6						■	■			
7							■	■		
8								■	■	
9									■	■
10	■									■

D. *Evaluasi Model*

Untuk mengevaluasi tingkat akurasi performansi model klasifikasi digunakan model evaluasi *Confusion Matrix* dan ROC *Curve*. *Confusion matrix* memberikan keputusan yang diperoleh dalam traning dan testing [1]. *Confusion matrix* memberikan penilaian performance klasifikasi berdasarkan objek dengan benar atau salah [3]. Confusion matrix merupakan matrik dua dimensi yang menggambarkan perbandingan antara hasil prediksi dengan kenyataan [11], ditunjukkan pada tabel 3 dibawah ini.

Confusion *matrix* merupakan tabel matrix yang terdiri dari dua kelas, yaitu kelas yang dianggap sebagai positif dan kelas yang dianggap sebagai negatif [8]. *Confusion matrix* berisi informasi aktual (actual) dan prediksi (*predicted*) pada model klasifikasi. *Confusion matrix* memberikan penilaian kinerja model klasifikasi berdasarkan jumlah objek yang diprediksi dengan benar dan salah [11].

TABEL IIIII
TABEL MODEL CONFUSION MATRIX

Classification	***Predicted Class***		
		Class = YES	***Class = NO***
Observed Class	***Class = YES***	**A** (*true positive*-TP)	**B** (*false negative*-FN)
	Class = NO	**C** (*false positive*-FP)	**D** (*true negative*-TN)

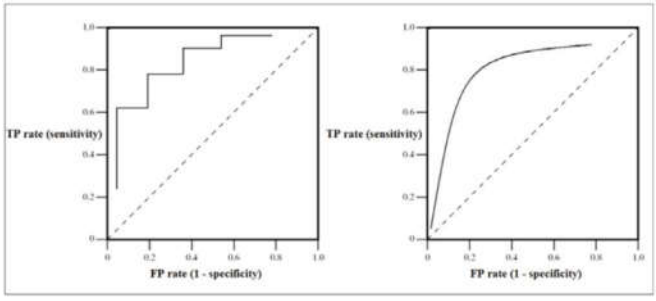
Keterangan:
True Positive(TP): Proporsi *positive* dalam *dataset* yang diklasifikasikan *positive*.
True Negative (TN): Proporsi *negative* dalam *dataset* yang diklasifikasikan *negative*.
False Positive (FP): Proporsi *positive* dalam *dataset* yang diklasifikasikan *negative*.
False Negative (FN): Proporsi *negative* dalam *dataset* yang diklasifikasikan *Positive*.

Untuk dapat melihat akurasi secara manual dilakukan perbandingan klasifikasi menggunakan *curva* ROC hasil ekspresi dari *confusion matrix*. Kurva ROC (*Receiver Operating Characteristic*) adalah cara lain untuk mengevaluasi akurasi dari klasifikasi secara visual [8]. Tingkat akurasi dapat di diagnosa sebagai berikut [3]:

Akurasi 0.90 – 1.00 = *Excellent classification*
Akurasi 0.80 – 0.90 = *Good classification*
Akurasi 0.70 – 0.80 = *Fair classification*
Akurasi 0.60 – 0.70 = *Poor classification*
Akurasi 0.50 – 0.60 = *Failure*

Untuk dapat melihat akurasi dapat dilakukan perbandingan klasifikasi menggunakan curva ROC hasil eksperesi dari confusion matrix. ROC menghasilkan dua garis dengan bentuk *true positives* sebagai garis vertikal dan *false positives* sebagai garis horizontal [8]. Kurva ROC adalah grafik antara sensitivitas (*true positive rate*) pada sumbu Y dengan 1-spesifisitas pada sumbu X (*false positive rate*), *curve* ROC ini seakan-akan menggambarkan tawar-menawar antara sumbu Y atau sensitivitas dengan sumbu X atau spesifisitas. Nilai dari kurva ROC ini diharapkan mempunyai nilai yang akurat dalam uji kuantitas dalam sebuah pengujian kasus. Menentukan nilai *cut off* pada uji *diagnostic* yang bersifat kontinyu dan membandingkan kualitas dari dua atau lebih uji *diagnostic.*

*A*rea *Under the ROC (Receiver Operating Characteristic) Curve* (AUROC atau AUC) adalah ukuran numerik untuk membedakan kinerja model, dan menunjukkan seberapa sukses dan benar peringkat model dengan memisahkan pengamatan positif dan negatif [13].



Gambar 2. Grafik ROC [3].

Pada Gambar 2 diatas garis diagonal membagi ruang ROC, yaitu: (a) poin diatas garis diagonal merupakan hasil klasifikasi yang baik, (b) point dibawah garis diagonal merupakan hasil klasifikasi yang buruk. Dapat disimpulkan bahwa, satu point pada kurva ROC adalah lebih baik dari pada yang lainnya jika arah garis melintang dari kiri bawah ke kanan atas didalam grafik. Ukuran AUC dihitung sebagai daerah dibawah kurva ROC dengan persamaan.

E. Komparasi Model

Dikarenakan hasil dari evaluasi *Confusion matrix* dan Kurva ROC menghasilkan data yang hampir sama sehingga menyulitkan dalam pengambilan keputusan maka dalam penelitian ini dilakukan uji beda. Untuk mengkomparasi hasil evaluasi digunakan model komparasi parametrik. Pada literatur ada dua jenis uji statistik, yaitu uji parametrik dan uji nonparametrik [12]. Uji parametrik yang dapat digunakan salah satunya adalah uji T (T-test). Pada penelitian ini akan digunakan model komparasi parametrik T-*Test*.

Uji t sampel berpasangan (*paired-samples t-test*) digunakan untuk membandingkan selisih dua rata-rata (*mean*) dari dua sampel yang berpasangan. Uji T berpasangan biasanya digunakan untuk menguji hipotesis penelitian apakah hasil setelah perbaikan lebih baik dari hasil sebelum perbaikan. Bentuk uji t sampel berpasangan akan menggunakan bentuk Uji t sampel berpasangan (*paired samples t-test*) dua sisi (*two-tailed*) dengan hipotesis mana yang diterima, ditentukan berdasarkan nilai P-value.

IV. HASIL DAN PEMBAHASAN

Eksperimen dilakukan pada laptop berbasis *Intel Core* i3 *Processor* dengan RAM 2 GB dan sistem operasi yang digunakan *windows* 8. Aplikasi yang digunakan untuk mengerjakan penelitian menggunakan *Rapid Miner Studio* 7.0

Pada tabel 4 dapat dilihat nilai akurasi dan AUC masing-masing algoritma berdasarkan model evaluasi *Confusion Matrix* dan UUC.

TABEL IVV
TABEL HASIL UJI BEDA T-TEST

	C45	NB	KNN	RI	RF
Accuracy	90.85%	82.52%	81.68%	90.57%	79.96%
AUC	0.904	0.891	0.500	0.906	0.800

Dari hasil evaluasi di atas langkah selanjutnya adalah dilakukan proses uji beda dengan menggunakan teknik uji beda parametrik *T-Test*. Hasil *T-test* dapat dilihat pada tabel 4. Pada Tabel 5 dapat terlihat bahwa algoritma *Rule Induction*, C4.5, dan *Random Forest* memiliki performa yang paling baik diantara algoritma yang lain.

TABEL V
TABEL UJI BEDA T-TEST

		C4.5	K-NN	NB	RI	RF
		0.908 +/- 0.047	0.817 +/- 0.076	0.825 +/- 0.052	0.906 +/- 0.066	0.800 +/- 0.096
C4.5	0.908 +/- 0.047		**0.004**	**0.001**	0.916	0.005
K-NN	0.817 +/- 0.076			0.775	**0.012**	0.661
NB	0.825 +/- 0.052				**0.007**	0.468
RI	0.906 +/- 0.066					**0.010**
RF	0.800 +/- 0.096					

V. KESIMPULAN

Berdasarkan penelitian perbandingan algoritma *Decision Tree, Naive Bayes*, K-NN, *Rule Induction*, dan *Random Forest* dengan *dataset* kelulusan mahasiswa diploma AMIK BSI "Pontianak", dapat disimpulkan bahwa dengan menggunakan uji beda *T-Test* algoritma *Rule Induction* dan C45 memiliki performa yang paling baik dalam memprediksi ketepatan kelulusan mahasiswa.

VI. UCAPAN TERIMA KASIH / ACKNOWLEDGMENT

Penulis mengucapkan terima kasih kepada segenap pihak yang telah membantu dalam pembuatan paper ini dan kepada tim jurnal JUSTIN yang telah meluangkan waktu untuk mereview dan menerbitkan paper in.

REFERENSI

[1] Bramer, M. "Principles of Data Mining". London: Springer-Verlag. 2006.

[2] Ernastuti, S. &. "Graduation Prediction of Gunadarma University Students Using Algorithm and Naive Bayes C4.5 Algoritmh". 2010.

[3] Gorunescu, F. "Data Mining Concepts Models and Techniques". Craiova: Springer. 2011.

[4] Han, & Kamber. "Data Mining Concepts and technique". San Francisco: Diane Cerra. 2006

[5] Hastuti, K. "Analisis Komparasi Algoritma Klasifikasi Data Mining V". Seminar Nasional Teknologi Informasi & Komunikasi Terapan(979 - 26 - 0255 - 0), 241249. Juni 2012.

[6] Kalyankar, Q. &. "Drop Out Feature of Student Data forAcademic Performance Using Decision Tree techniques". Global Journal of Computer Science and Technology, 2-4. 2010.

[7] Tahyudin, I. " Comparing Clasification Algorithm Of Data Mining to Predict the Graduation Students on Time". Information Systems International Conference (ISICO). Desember 2013.

[8] Vercellis. "Business Intelligence: Data Mining and Optimization for Decision Making Decision Making". John Willey & Sons Inc: Southern Gate. 2009.

[9] Vrettos, K. &. " Sentivity Analysis of Neural Network for Identifying the Factors for Collage Students Success". World Congress on Computer Science and Information Engineering. (978-0-7695-3507-4). 2009.

[10] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining: Practical Machine Learning Tools and Techniques (The Morgan Kaufmann series in data management systems). Morgan Kaufmann Publishers is an imprint of Elsevier. Burlington. 2011.

[11] Gorunescu, F. Data mining: concepts and techniques. Springer-Verlag. Berlin. 2011.

[12] García, S., Fernández, A., Luengo, J., & Herrera, F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences, 180(10), 2044–2064. 2010

[13] Attenberg, J., & Ertekin, S. (2013). Class Imbalance and Active Learning. In H. He, & Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*, New Jersey: John Wiley & Sons. pp. 101-149