

# Machine Learning in Medicine - a Complete Overview



Ton J. Cleophas • Aeilko H. Zwinderman

# Machine Learning in Medicine - a Complete Overview

With the help from HENNY I. CLEOPHAS-ALLERS,  
BChem



Ton J. Cleophas  
Department Medicine  
Albert Schweitzer Hospital  
Sliedrecht, The Netherlands

Aeilko H. Zwinderman  
Department Biostatistics and Epidemiology  
Academic Medical Center  
Amsterdam, The Netherlands

Additional material to this book can be downloaded from <http://extras.springer.com>.

ISBN 978-3-319-15194-6                      ISBN 978-3-319-15195-3 (eBook)  
DOI 10.1007/978-3-319-15195-3

Library of Congress Control Number: 2015930334

Springer Cham Heidelberg New York Dordrecht London  
© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Preface

The amount of data stored in the world's databases doubles every 20 months, as estimated by Usama Fayyad, one of the founders of machine learning and co-author of the book *Advances in Knowledge Discovery and Data Mining* (ed. by the American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996), and clinicians, familiar with traditional statistical methods, are at a loss to analyze them.

Traditional methods have, indeed, difficulty to identify outliers in large datasets, and to find patterns in big data and data with multiple exposure/outcome variables. In addition, analysis-rules for surveys and questionnaires, which are currently common methods of data collection, are, essentially, missing. Fortunately, the new discipline, machine learning, is able to cover all of these limitations.

So far, medical professionals have been rather reluctant to use machine learning. Ravinda Khattree, co-author of the book *Computational Methods in Biomedical Research* (ed. by Chapman & Hall, Baton Rouge, LA, USA, 2007) suggests that there may be historical reasons: technological (doctors are better than computers (?)), legal, cultural (doctors are better trusted). Also, in the field of diagnosis making, few doctors may want a computer checking them, are interested in collaboration with a computer or with computer engineers.

Adequate health and health care will, however, soon be impossible without proper data supervision from modern machine learning methodologies like cluster models, neural networks, and other data mining methodologies. The current book is the first publication of a complete overview of machine learning methodologies for the medical and health sector, and it was written as a training companion, and as a must-read, not only for physicians and students, but also for anyone involved in the process and progress of health and health care.

Some of the 80 chapters have already appeared in Springer's Cookbook Briefs, but they have been rewritten and updated. All of the chapters have two core characteristics. First, they are intended for current usage, and they are, particularly, concerned with improving that usage. Second, they try and tell what readers need to know in order to understand the methods.

In a nonmathematical way, stepwise analyses of the below three most important classes of machine learning methods will be reviewed:

Cluster and classification models (Chaps. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18),

(Log)linear models (Chaps. 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, and 49),

Rules models (Chaps. 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, and 80).

The book will include basic methodologies like typology of medical data, quantile-quantile plots for making a start with your data, rate analysis and trend analysis as more powerful alternatives to risk analysis and traditional tests, probit models for binary effects on treatment frequencies, higher order polynomes for circadian phenomena, contingency tables and its myriad applications. Particularly, Chaps. 9, 14, 15, 18, 45, 48, 49, 79, and 80 will review these methodologies.

Chapter 7 describes the use of visualization processes instead of calculus methods for data mining. Chapter 8 describes the use of trained clusters, a scientifically more appropriate alternative to traditional cluster analysis. Chapter 69 describes evolutionary operations (evops), and the evop calculators, already widely used for chemical and technical process improvement.

Various automated analyses and simulation models are in Chaps. 4, 29, 31, and 32. Chapters 67, 70, 71 review spectral plots, Bayesian networks, and support vector machines. A first description of several methods already employed by technical and market scientists, and of their suitabilities for clinical research, is given in Chaps. 37, 38, 39, and 56 (ordinal scalings for inconsistent intervals, loglinear models for varying incident risks, and iteration methods for cross-validations).

Modern methodologies like interval censored analyses, exploratory analyses using pivoting trays, repeated measures logistic regression, doubly multivariate analyses for health assessments, and gamma regression for best fit prediction of health parameters are reviewed in Chaps. 10, 11, 12, 13, 16, 17, 42, 46, and 47.

In order for the readers to perform their own analyses, SPSS data files of the examples are given in [extras.springer.com](http://extras.springer.com), as well as XML (eXtended Markup Language), SPS (Syntax), and ZIP (compressed) files for outcome predictions in future patients. Furthermore, four csv type excel files are available for data analysis in the Konstanz information miner (Knime) and Weka (Waikato University New Zealand) miner, widely approved free machine learning software packages on the internet since 2006. Also a first introduction is given to SPSS modeler (SPSS' data mining workbench, Chaps. 61, 64, 65), and to SPSS Amos, the graphical and non-graphical data analyzer for the identification of cause-effect relationships as principle goal of research (Chaps. 48 and 49). The free Davidwees polynomial grapher is used in Chap. 79.

This book will demonstrate that machine learning performs sometimes better than traditional statistics does. For example, if the data perfectly fit the cut-offs for node splitting, because, e.g.,  $\text{ages} > 55$  years give an exponential rise in infarctions, then decision trees, optimal binning, and optimal scaling will be better

analysis-methods than traditional regression methods with age as continuous predictor. Machine learning may have little options for adjusting confounding and interaction, but you can add propensity scores and interaction variables to almost any machine learning method.

Each chapter will start with purposes and scientific questions. Then, step-by-step analyses, using both real data and simulated data examples, will be given. Finally, a paragraph with conclusion, and references to the corresponding sites of three introductory textbooks previously written by the same authors, is given.

Lyon, France  
December 2015

Ton J. Cleophas  
Aeilko H. Zwinderman



# Contents

## Part I Cluster and Classification Models

<b>1</b>	<b>Hierarchical Clustering and K-Means Clustering to Identify Subgroups in Surveys (50 Patients)</b> .....	3
	General Purpose .....	3
	Specific Scientific Question .....	3
	Hierarchical Cluster Analysis.....	4
	K-Means Cluster Analysis.....	6
	Conclusion.....	7
	Note .....	8
<b>2</b>	<b>Density-Based Clustering to Identify Outlier Groups in Otherwise Homogeneous Data (50 Patients)</b> .....	9
	General Purpose .....	9
	Specific Scientific Question .....	9
	Density-Based Cluster Analysis .....	10
	Conclusion.....	11
	Note .....	11
<b>3</b>	<b>Two Step Clustering to Identify Subgroups and Predict Subgroup Memberships in Individual Future Patients (120 Patients)</b> .....	13
	General Purpose .....	13
	Specific Scientific Question .....	13
	The Computer Teaches Itself to Make Predictions .....	14
	Conclusion.....	15
	Note .....	15
<b>4</b>	<b>Nearest Neighbors for Classifying New Medicines (2 New and 25 Old Opioids)</b> .....	17
	General Purpose .....	17
	Specific Scientific Question .....	17

Example.....	17
Conclusion.....	24
Note.....	24
<b>5 Predicting High-Risk-Bin Memberships (1,445 Families).....</b>	<b>25</b>
General Purpose .....	25
Specific Scientific Question .....	25
Example.....	25
Optimal Binning.....	26
Conclusion.....	29
Note.....	29
<b>6 Predicting Outlier Memberships (2,000 Patients).....</b>	<b>31</b>
General Purpose .....	31
Specific Scientific Question .....	31
Example.....	31
Conclusion.....	34
Note.....	34
<b>7 Data Mining for Visualization of Health Processes (150 Patients).....</b>	<b>35</b>
General Purpose .....	35
Primary Scientific Question .....	35
Example.....	36
Knime Data Miner.....	37
Knime Workflow .....	38
Box and Whiskers Plots .....	39
Lift Chart.....	39
Histogram.....	40
Line Plot.....	41
Matrix of Scatter Plots .....	42
Parallel Coordinates .....	43
Hierarchical Cluster Analysis with SOTA (Self Organizing Tree Algorithm) .....	44
Conclusion.....	45
Note.....	46
<b>8 Trained Decision Trees for a More Meaningful Accuracy   (150 Patients) .....</b>	<b>47</b>
General Purpose .....	47
Primary Scientific Question .....	47
Example.....	48
Downloading the Knime Data Miner .....	49
Knime Workflow .....	50
Conclusion.....	52
Note.....	52

- 9 Typology of Medical Data (51 Patients) ..... 53**
  - General Purpose ..... 53
  - Primary Scientific Question ..... 54
  - Example..... 54
    - Nominal Variable..... 55
    - Ordinal Variable..... 56
    - Scale Variable ..... 57
  - Conclusion..... 59
  - Note..... 60
- 10 Predictions from Nominal Clinical Data (450 Patients) ..... 61**
  - General Purpose ..... 61
  - Primary Scientific Question ..... 61
  - Example..... 61
  - Conclusion..... 65
  - Note..... 65
- 11 Predictions from Ordinal Clinical Data (450 Patients)..... 67**
  - General Purpose ..... 67
  - Primary Scientific Question ..... 67
  - Example..... 68
  - Conclusion..... 70
  - Note..... 70
- 12 Assessing Relative Health Risks (3,000 Subjects)..... 71**
  - General Purpose ..... 71
  - Primary Scientific Question ..... 71
  - Example..... 71
  - Conclusion..... 75
  - Note..... 75
- 13 Measuring Agreement (30 Patients) ..... 77**
  - General Purpose ..... 77
  - Primary Scientific Question ..... 77
  - Example..... 77
  - Conclusion..... 79
  - Note..... 79
- 14 Column Proportions for Testing Differences Between Outcome Scores (450 Patients)..... 81**
  - General Purpose ..... 81
  - Specific Scientific Question ..... 81
  - Example..... 81
  - Conclusion..... 85
  - Note..... 85

**15 Pivoting Trays and Tables for Improved Analysis of Multidimensional Data (450 Patients)**..... 87

    General Purpose ..... 87

    Primary Scientific Question ..... 87

    Example..... 87

    Conclusion..... 94

    Note ..... 94

**16 Online Analytical Procedure Cubes, a More Rapid Approach to Analyzing Frequencies (450 Patients)** ..... 95

    General Purpose ..... 95

    Primary Scientific Question ..... 95

    Example..... 95

    Conclusion..... 99

    Note ..... 99

**17 Restructure Data Wizard for Data Classified the Wrong Way (20 Patients)** ..... 101

    General Purpose ..... 101

    Primary Scientific Question ..... 103

    Example..... 103

    Conclusion..... 104

    Note ..... 104

**18 Control Charts for Quality Control of Medicines (164 Tablet Desintegration Times)** ..... 105

    General Purpose ..... 105

    Primary Scientific Question ..... 105

    Example..... 106

    Conclusion..... 109

    Note ..... 110

**Part II (Log) Linear Models**

**19 Linear, Logistic, and Cox Regression for Outcome Prediction with Unpaired Data (20, 55, and 60 Patients)** ..... 113

    General Purpose ..... 113

    Specific Scientific Question ..... 113

    Linear Regression, the Computer Teaches Itself to Make Predictions..... 114

    Conclusion..... 116

    Note ..... 116

    Logistic Regression, the Computer Teaches Itself to Make Predictions ... 116

    Conclusion..... 118

    Note ..... 118

    Cox Regression, the Computer Teaches Itself to Make Predictions ..... 118

    Conclusion..... 121

    Note ..... 121

**20 Generalized Linear Models for Outcome Prediction with Paired Data (100 Patients and 139 Physicians)** ..... 123

    General Purpose ..... 123

    Specific Scientific Question ..... 123

    Generalized Linear Modeling, the Computer Teaches Itself to Make Predictions ..... 123

    Conclusion..... 125

    Generalized Estimation Equations, the Computer Teaches Itself to Make Predictions ..... 126

    Conclusion..... 129

    Note ..... 129

**21 Generalized Linear Models Event-Rates (50 Patients)**..... 131

    General Purpose ..... 131

    Specific Scientific Question ..... 131

    Example..... 131

    The Computer Teaches Itself to Make Predictions ..... 132

    Conclusion..... 135

    Note ..... 135

**22 Factor Analysis and Partial Least Squares (PLS) for Complex-Data Reduction (250 Patients)** ..... 137

    General Purpose ..... 137

    Specific Scientific Question ..... 137

    Factor Analysis..... 138

    Partial Least Squares Analysis (PLS)..... 140

    Traditional Linear Regression..... 142

    Conclusion..... 142

    Note ..... 142

**23 Optimal Scaling of High-Sensitivity Analysis of Health Predictors (250 Patients)**..... 143

    General Purpose ..... 143

    Specific Scientific Question ..... 143

    Traditional Multiple Linear Regression ..... 144

    Optimal Scaling Without Regularization ..... 145

    Optimal Scaling With Ridge Regression..... 146

    Optimal Scaling With Lasso Regression..... 147

    Optimal Scaling With Elastic Net Regression..... 147

    Conclusion..... 148

    Note ..... 148

**24 Discriminant Analysis for Making a Diagnosis from Multiple Outcomes (45 Patients)** ..... 149

    General Purpose ..... 149

    Specific Scientific Question ..... 149

    The Computer Teaches Itself to Make Predictions ..... 150

    Conclusion..... 153

    Note ..... 153

**25 Weighted Least Squares for Adjusting Efficacy Data with Inconsistent Spread (78 Patients)**..... 155

    General Purpose ..... 155

    Specific Scientific Question ..... 155

    Weighted Least Squares ..... 156

    Conclusion..... 158

    Note ..... 158

**26 Partial Correlations for Removing Interaction Effects from Efficacy Data (64 Patients)**..... 159

    General Purpose ..... 159

    Specific Scientific Question ..... 159

    Partial Correlations..... 160

    Conclusion..... 162

    Note ..... 163

**27 Canonical Regression for Overall Statistics of Multivariate Data (250 Patients)** ..... 165

    General Purpose ..... 165

    Specific Scientific Question ..... 165

    Canonical Regression..... 166

    Conclusion..... 169

    Note ..... 169

**28 Multinomial Regression for Outcome Categories (55 Patients)**..... 171

    General Purpose ..... 171

    Specific Scientific Question ..... 171

    The Computer Teaches Itself to Make Predictions ..... 172

    Conclusion..... 174

    Note ..... 174

**29 Various Methods for Analyzing Predictor Categories (60 and 30 Patients)**..... 175

    General Purpose ..... 175

    Specific Scientific Questions..... 175

    Example 1..... 175

    Example 2..... 179

    Conclusion..... 182

    Note ..... 182

**30 Random Intercept Models for Both Outcome and Predictor Categories (55 patients)**..... 183

    General Purpose ..... 183

    Specific Scientific Question ..... 184

    Example..... 184

    Conclusion..... 187

    Note ..... 187

<b>31 Automatic Regression for Maximizing Linear Relationships (55 patients)</b> .....	189
General Purpose .....	189
Specific Scientific Question .....	189
Data Example .....	189
The Computer Teaches Itself to Make Predictions .....	192
Conclusion.....	193
Note.....	194
<b>32 Simulation Models for Varying Predictors (9,000 Patients)</b> .....	195
General Purpose .....	195
Specific Scientific Question .....	195
Instead of Traditional Means and Standard Deviations, Monte Carlo Simulations of the Input and Outcome Variables are Used to Model the Data. This Enhances Precision, Particularly, With non-Normal Data .....	196
Conclusion.....	200
Note.....	201
<b>33 Generalized Linear Mixed Models for Outcome Prediction from Mixed Data (20 Patients)</b> .....	203
General Purpose .....	203
Specific Scientific Question .....	203
Example.....	203
Conclusion.....	206
Note.....	206
<b>34 Two-Stage Least Squares (35 Patients)</b> .....	207
General Purpose .....	207
Primary Scientific Question .....	207
Example.....	208
Conclusion.....	210
Note.....	210
<b>35 Autoregressive Models for Longitudinal Data (120 Mean Monthly Population Records)</b> .....	211
General Purpose .....	211
Specific Scientific Question .....	211
Example.....	212
Conclusion.....	216
Note.....	217
<b>36 Variance Components for Assessing the Magnitude of Random Effects (40 Patients)</b> .....	219
General Purpose .....	219
Primary Scientific Question .....	219
Example.....	220
Conclusion.....	222
Note.....	222

**37 Ordinal Scaling for Clinical Scores with Inconsistent Intervals (900 Patients)** ..... 223

    General Purpose ..... 223

    Primary Scientific Questions ..... 223

    Example ..... 223

    Conclusion ..... 227

    Note ..... 227

**38 Loglinear Models for Assessing Incident Rates with Varying Incident Risks (12 Populations)** ..... 229

    General Purpose ..... 229

    Primary Scientific Question ..... 230

    Example ..... 230

    Conclusion ..... 232

    Note ..... 232

**39 Loglinear Modeling for Outcome Categories (445 Patients)** ..... 233

    General Purpose ..... 233

    Primary Scientific Question ..... 233

    Example ..... 234

    Conclusion ..... 239

    Note ..... 239

**40 Heterogeneity in Clinical Research: Mechanisms Responsible (20 Studies)** ..... 241

    General Purpose ..... 241

    Primary Scientific Question ..... 241

    Example ..... 242

    Conclusion ..... 244

    Note ..... 244

**41 Performance Evaluation of Novel Diagnostic Tests (650 and 588 Patients)** ..... 245

    General Purpose ..... 245

    Primary Scientific Question ..... 245

    Example ..... 245

        Binary Logistic Regression ..... 248

        C-Statistics ..... 249

    Conclusion ..... 251

    Note ..... 251

**42 Quantile-Quantile Plots, a Good Start for Looking at Your Medical Data (50 Cholesterol Measurements and 58 Patients)** ..... 253

    General Purpose ..... 253

    Specific Scientific Question ..... 253

    Q-Q Plots for Assessing Departures from Normality ..... 253

Q-Q Plots as Diagnostics for Fitting Data to Normal (and Other Theoretical) Distributions.....	256
Conclusion.....	258
Note.....	259
<b>43 Rate Analysis of Medical Data Better than Risk Analysis (52 Patients) .....</b>	<b>261</b>
General Purpose .....	261
Specific Scientific Question .....	261
Example.....	261
Conclusion.....	264
Note.....	264
<b>44 Trend Tests Will Be Statistically Significant if Traditional Tests Are Not (30 and 106 Patients) .....</b>	<b>265</b>
General Purpose .....	265
Specific Scientific Questions.....	265
Example 1.....	265
Example 2.....	267
Conclusion.....	269
Note.....	269
<b>45 Doubly Multivariate Analysis of Variance for Multiple Observations from Multiple Outcome Variables (16 Patients) .....</b>	<b>271</b>
General Purpose .....	271
Primary Scientific Question .....	271
Example.....	272
Conclusion.....	276
Note.....	276
<b>46 Probit Models for Estimating Effective Pharmacological Treatment Dosages (14 Tests) .....</b>	<b>279</b>
General Purpose .....	279
Primary Scientific Question .....	279
Example.....	279
Simple Probit Regression .....	279
Multiple Probit Regression.....	282
Conclusion.....	286
Note.....	287
<b>47 Interval Censored Data Analysis for Assessing Mean Time to Cancer Relapse (51 Patients) .....</b>	<b>289</b>
General Purpose .....	289
Primary Scientific Question .....	289
Example.....	290
Conclusion.....	292
Note.....	293

**48 Structural Equation Modeling (SEM) with SPSS Analysis of Moment Structures (Amos) for Cause Effect Relationships I (35 Patients)**..... 295

    General Purpose ..... 295

    Primary Scientific Question ..... 296

    Example..... 296

    Conclusion..... 300

    Note..... 300

**49 Structural Equation Modeling (SEM) with SPSS Analysis of Moment Structures (Amos) for Cause Effect Relationships in Pharmacodynamic Studies II (35 Patients)** ..... 301

    General Purpose ..... 301

    Primary Scientific Question ..... 302

    Example..... 302

    Conclusion..... 306

    Note..... 306

**Part III Rules Models**

**50 Neural Networks for Assessing Relationships That Are Typically Nonlinear (90 Patients)** ..... 309

    General Purpose ..... 309

    Specific Scientific Question ..... 309

    The Computer Teaches Itself to Make Predictions ..... 310

    Conclusion..... 311

    Note..... 312

**51 Complex Samples Methodologies for Unbiased Sampling (9,678 Persons)**..... 313

    General Purpose ..... 313

    Specific Scientific Question ..... 313

    The Computer Teaches Itself to Predict Current Health Scores from Previous Health Scores ..... 315

    The Computer Teaches Itself to Predict Individual Odds Ratios of Current Health Scores Versus Previous Health Scores..... 317

    Conclusion..... 318

    Note..... 319

**52 Correspondence Analysis for Identifying the Best of Multiple Treatments in Multiple Groups (217 Patients)** ..... 321

    General Purpose ..... 321

    Specific Scientific Question ..... 321

    Correspondence Analysis..... 322

    Conclusion..... 325

    Note..... 325

**53 Decision Trees for Decision Analysis (1,004 and 953 Patients)..... 327**  
 General Purpose ..... 327  
 Specific Scientific Question ..... 327  
 Decision Trees with a Binary Outcome ..... 327  
 Decision Trees with a Continuous Outcome ..... 331  
 Conclusion..... 334  
 Note ..... 334

**54 Multidimensional Scaling for Visualizing Experienced Drug Efficacies (14 Pain-Killers and 42 Patients) ..... 335**  
 General Purpose ..... 335  
 Specific Scientific Question ..... 335  
 Proximity Scaling..... 336  
 Preference Scaling..... 338  
 Conclusion..... 343  
 Note ..... 344

**55 Stochastic Processes for Long Term Predictions from Short Term Observations ..... 345**  
 General Purpose ..... 345  
 Specific Scientific Questions..... 345  
 Example 1..... 345  
 Example 2..... 347  
 Example 3..... 349  
 Conclusion..... 351  
 Note ..... 351

**56 Optimal Binning for Finding High Risk Cut-offs (1,445 Families)..... 353**  
 General Purpose ..... 353  
 Specific Scientific Question ..... 353  
 Optimal Binning..... 354  
 Conclusion..... 357  
 Note ..... 357

**57 Conjoint Analysis for Determining the Most Appreciated Properties of Medicines to Be Developed (15 Physicians) ..... 359**  
 General Purpose ..... 359  
 Specific Scientific Question ..... 359  
 Constructing an Analysis Plan ..... 359  
 Performing the Final Analysis..... 361  
 Conclusion..... 364  
 Note ..... 364

**58 Item Response Modeling for Analyzing Quality of Life with Better Precision (1,000 Patients) ..... 365**

General Purpose ..... 365

Primary Scientific Question ..... 365

Example..... 365

Conclusion..... 369

Note ..... 369

**59 Survival Studies with Varying Risks of Dying (50 and 60 Patients)..... 371**

General Purpose ..... 371

Primary Scientific Question ..... 371

Examples ..... 371

    Cox Regression with a Time-Dependent Predictor ..... 371

    Cox Regression with a Segmented Time-Dependent Predictor ..... 373

Conclusion..... 374

Note ..... 375

**60 Fuzzy Logic for Improved Precision of Dose-Response Data (8 Induction Dosages) ..... 377**

General Purpose ..... 377

Specific Scientific Question ..... 377

Example..... 378

Conclusion..... 381

Note ..... 381

**61 Automatic Data Mining for the Best Treatment of a Disease (90 Patients) ..... 383**

General Purpose ..... 383

Specific Scientific Question ..... 383

Example..... 383

Step 1 Open SPSS Modeler..... 385

Step 2 The Distribution Node..... 385

Step 3 The Data Audit Node ..... 386

Step 4 The Plot Node ..... 387

Step 5 The Web Node..... 388

Step 6 The Type and c5.0 Nodes ..... 389

Step 7 The Output Node..... 390

Conclusion..... 390

Note ..... 390

**62 Pareto Charts for Identifying the Main Factors of Multifactorial Outcomes (2,000 Admissions to Hospital)..... 391**

General Purpose ..... 391

Primary Scientific Question ..... 391

Example..... 392

Conclusion..... 396

Note ..... 396

<b>63 Radial Basis Neural Networks for Multidimensional Gaussian Data (90 Persons)</b> .....	397
General Purpose .....	397
Specific Scientific Question .....	397
Example.....	397
The Computer Teaches Itself to Make Predictions .....	398
Conclusion.....	400
Note.....	400
<b>64 Automatic Modeling of Drug Efficacy Prediction (250 Patients)</b> .....	401
General Purpose .....	401
Specific Scientific Question .....	401
Example.....	401
Step 1 Open SPSS Modeler (14.2).....	402
Step 2 The Statistics File Node .....	403
Step 3 The Type Node .....	403
Step 4 The Auto Numeric Node .....	404
Step 5 The Expert Node .....	405
Step 6 The Settings Tab.....	407
Step 7 The Analysis Node .....	407
Conclusion.....	408
Note.....	408
<b>65 Automatic Modeling for Clinical Event Prediction (200 Patients)</b> ....	409
General Purpose .....	409
Specific Scientific Question .....	409
Example.....	409
Step 1 Open SPSS Modeler (14.2).....	410
Step 2 The Statistics File Node .....	411
Step 3 The Type Node .....	411
Step 4 The Auto Classifier Node .....	412
Step 5 The Expert Tab .....	413
Step 6 The Settings Tab.....	414
Step 7 The Analysis Node .....	415
Conclusion.....	416
Note.....	416
<b>66 Automatic Newton Modeling in Clinical Pharmacology (15 Alfentanil Dosages, 15 Quinidine Time-Concentration Relationships)</b> .....	417
General Purpose .....	417
Specific Scientific Question .....	418
Examples .....	418
Dose-Effectiveness Study.....	418
Time-Concentration Study .....	420
Conclusion.....	422
Note.....	422

**67 Spectral Plots for High Sensitivity Assessment of Periodicity (6 Years’ Monthly C Reactive Protein Levels)** ..... 423

    General Purpose ..... 423

    Specific Scientific Question ..... 423

    Example..... 423

    Conclusion..... 427

    Note ..... 428

**68 Runs Test for Identifying Best Regression Models (21 Estimates of Quantity and Quality of Patient Care)**..... 429

    General Purpose ..... 429

    Primary Scientific Question ..... 429

    Example..... 429

    Conclusion..... 433

    Note ..... 433

**69 Evolutionary Operations for Process Improvement (8 Operation Room Air Condition Settings)** ..... 435

    General Purpose ..... 435

    Specific Scientific Question ..... 435

    Example..... 436

    Conclusion..... 437

    Note ..... 438

**70 Bayesian Networks for Cause Effect Modeling (600 Patients)**..... 439

    General Purpose ..... 439

    Primary Scientific Question ..... 439

    Example..... 440

    Binary Logistic Regression in SPSS ..... 440

    Konstanz Information Miner (Knime) ..... 441

    Knime Workflow ..... 442

    Conclusion..... 443

    Note ..... 444

**71 Support Vector Machines for Imperfect Nonlinear Data (200 Patients with Sepsis)** ..... 445

    General Purpose ..... 445

    Primary Scientific Question ..... 445

    Example..... 446

    Knime Data Miner..... 446

    Knime Workflow ..... 447

    File Reader Node..... 447

    The Nodes X-Partitioner, svm Learner, svm Predictor, X-Aggregator..... 448

    Error Rates ..... 448

    Prediction Table..... 448

    Conclusion..... 449

    Note ..... 449

- 72 Multiple Response Sets for Visualizing Clinical Data Trends (811 Patient Visits)**..... 451
  - General Purpose ..... 451
  - Specific Scientific Question ..... 451
  - Example..... 451
  - Conclusion..... 457
  - Note ..... 457
  
- 73 Protein and DNA Sequence Mining**..... 459
  - General Purpose ..... 459
  - Specific Scientific Question ..... 459
  - Data Base Systems on the Internet..... 460
  - Example 1..... 461
  - Example 2..... 462
  - Example 3..... 462
  - Example 4..... 463
  - Conclusion..... 464
  - Note ..... 464
  
- 74 Iteration Methods for Crossvalidations (150 Patients with Pneumonia)**..... 465
  - General Purpose ..... 465
  - Primary Scientific Question ..... 465
  - Example..... 465
  - Downloading the Knime Data Miner ..... 466
  - Knime Workflow ..... 467
  - Crossvalidation..... 467
  - Conclusion..... 469
  - Note ..... 469
  
- 75 Testing Parallel-Groups with Different Sample Sizes and Variances (5 Parallel-Group Studies)**..... 471
  - General Purpose ..... 471
  - Primary Scientific Question ..... 471
  - Examples ..... 472
  - Conclusion..... 473
  - Note ..... 473
  
- 76 Association Rules Between Exposure and Outcome (50 and 60 Patients)**..... 475
  - General Purpose ..... 475
  - Primary Scientific Question ..... 475
  - Example..... 475
    - Example One ..... 477
    - Example Two..... 478
  - Conclusion..... 479
  - Note ..... 479

**77 Confidence Intervals for Proportions and Differences in Proportions (100 and 75 Patients) ..... 481**  
General Purpose ..... 481  
Primary Scientific Question ..... 481  
Example..... 482  
    Confidence Intervals of Proportions..... 482  
    Confidence Intervals of Differences in Proportions ..... 483  
Conclusion..... 484  
Note..... 484

**78 Ratio Statistics for Efficacy Analysis of New Drugs (50 Patients) ..... 485**  
General Purpose ..... 485  
Primary Scientific Question ..... 485  
Example..... 485  
Conclusion..... 489  
Note..... 489

**79 Fifth Order Polynomes of Circadian Rhythms (1 Patient with Hypertension) ..... 491**  
General Purpose ..... 491  
Primary Scientific Question ..... 492  
Example..... 492  
Conclusion..... 495  
Note..... 496

**80 Gamma Distribution for Estimating the Predictors of Medical Outcome Scores (110 Patients)..... 497**  
General Purpose ..... 497  
Primary Scientific Question ..... 498  
Example..... 498  
Conclusion..... 503  
Note..... 503

**Index ..... 505**