

**DETERMINING THE EFFECT OF CURRICULUM AND
FACILITIES ON ACADEMIC ACHIEVEMENT USING DATA
MINING APPROACH**

**A thesis submitted to the Faculty of Information Technology in partial
Fulfillment of the requirement for the degree
Master of Science (Intelligent System)
Universiti Utara Malaysia**

**By
SURENAH SEDEK**



KOLEJ SASTERA DAN SAINS
(College of Arts and Sciences)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certify that)

SURENAH SEDEK

calon untuk Ijazah
(candidate for the degree of) **MSc. (Intelligent System)**

telah mengemukakan kertas projek yang bertajuk
(has presented his/her project paper of the following title)

DETERMINING THE EFFECTS OF CURRICULUM AND FACILITIES
ON ACADEMIC ACHIEVEMENT USING DATA MINING

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan
dan meliputi bidang ilmu dengan memuaskan.
(that the project paper acceptable in form and content, and that a satisfactory
knowledge of the field is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor): **ASSOC. PROF. FADZILAH SIRAJ**

Tandatangan
(Signature)

:

Tarikh (Date) :

17/06/08

Nama Penyelia Kedua
(Name of 2nd Supervisor): **MISS NOORAINI YUSOFF**

Tandatangan
(Signature)

:

Tarikh (Date) :

17/06/08

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor, in her absence, by the Dean of the Faculty of Information Technology. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain should not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Request for permission to copy or to make use of material in this thesis, in whole or in part should be addressed to:

Dean of the Faculty of Information Technology

Universiti Utara Malaysia

06010 UUM Sintok

Kedah Darul Aman

ABSTRAK

(BAHASA MELAYU)

Bidang pendidikan mempunyai pelbagai aplikasi perlombongan data yang menarik dan mencabar, serta dikenalpasti sebagai satu teknik yang berpotensi digunakan untuk membantu tenaga pengajar dan pelajar, dan memperbaiki kualiti sistem pendidikan. Berdasarkan masalah yang telah ditemui teknik perlombongan data dengan mengaplikasikan moden statistik dan teknologi pengkomputeran dapat menyelesaikan masalah pencarian maklumat yang penting di dalam pangkalan data yang besar. Kajian ini telah mengaplikasikan teknik perlombongan data dengan tujuan untuk mengkaji maklumat yang tersembunyi dan penting yang boleh menjejaskan pencapaian akademik dikalangan responden. Responden yang terlibat didalam kajian ini merupakan responden dari universiti awam yang telah menamatkan pengajian pada tahun 2007. Soal selidik yang digunakan untuk kajian ini berdasarkan soal selidik Kajian Pengesanan Graduan Kementerian Pengajian Tinggi dan kajian ini hanya menganalisa soalan di bahagian latarbelakang responden, kurikulum dan kemudahan yang telah disediakan di setiap institusi. Objektif utama kajian ini adalah untuk mengkaji sama ada kurikulum dan kemudahan yang telah disediakan di setiap institusi mempengaruhi keputusan akademik pelajar atau tidak. Bilangan responden yang terlibat didalam kajian ini adalah seramai 55,315 orang untuk bahagian pendekatan penerangan data dan 39,801 untuk bahagian pendekatan ramalan. Pendekatan ramalan akan menggunakan dua kaedah yang berbeza iaitu kaedah statistik yang menggunakan perisian "Statistical Packages for Sosial Science (SPSS)" dan kaedah rangkaian neural yang menggunakan perisian "Statistical Analytical Software (SAS)". Setelah semua ujian untuk melihat kesan kurikulum dan kemudahan di IPTA menggunakan kedua-dua kaedah selesai. Keputusan yang diperolehi menunjukkan rangkaian neural memberikan peratusan tertinggi berbanding statistik. Berdasarkan keputusan yang telah diperolehi menunjukkan bahawa rangkaian neural memperoleh peratusan tertinggi iaitu 89.47 % dengan

mengukur item kurikulum dan latarbelakang dengan pencapaian akademik responden. Secara keseluruhannya, model rangkaian neural memperoleh peratusan lebih daripada 80 % manakala statistik hanya menghasilkan kurang daripada 50%.

ABSTRACT (ENGLISH)

Education domain offers many interest and challenge in data mining applications that potentially identified as a tool to help both educators and students to improve the quality of education system. Data Mining applies modern statistical and computational technologies to the problem of finding useful hidden patterns within large databases. Thus, this study applied data mining technique to identify the hidden information that affects the academic achievement among respondents. The respondents for this study are consists of all public university students which complete their study within year 2007. The questionnaire that has been used in this study was adopted from Kajian Pengesanan Graduan, Kementerian Pengajian Tinggi and it focuses on curriculum and facilities that have been provided by universities. The aims of this study is to determine whether the curriculum and facilities that provided by university has an effect on students academic achievement. 55,315 respondents data were used for descriptive task while 39,801 data for predictive task. Both data mining approaches, namely the descriptive and predictive have been utilized to perform the analysis prior to build the model. For descriptive purposes, frequency, cross tabulation and correlation coefficients were computed to check whether significant correlation exists. For predictive modeling, logistic regression and neural network were used. Statistical Packages for Sosial Science (SPSS) was used for regression technique and Statistical Analytical Software (SAS) for Neural Network modeling. Then, the online questionnaire was integrated with Neural Network model to predict future student academic achievement. The findings in this study suggest neural network is the best model compared to logistic regression to measure the effect of curriculum and facilities on academic achievement. The highest accuracy from neural network is 89.47%, when demographics and curriculum become the contributing variables to academic achievement. Most of the neural network model accuracy is over than 80% while logistic regression accuracy is below than 50 %.

ACKNOWLEDGEMENT

In the name of Allah, Most Gracious, Most Merciful. Peace upon the prophet, Muhammad S.A.W, a foremost praise and thankful to Allah for His blessing, giving me the strength in completing this research.

On the other hand, my special thanks to Universiti Utara Malaysia (UUM) for facilities and resource that has been provided to all students. With all the facilities that have been provided, it helped me to make my research completed and succeeded.

I would like to extend my thanks and gratitude to my supervisor Associate Professor Fadzilah Siraj and Miss Nooraini Yusoff for the guidance, patience, encouragements, advice and flourish of knowledge during completing this research.

My warm appreciation dedicates to the Ministry of Higher Education who has gives their cooperation during this research.

Last but not least, a lasting heartfelt to my family and all my friends for all of the love and support in completing this course, as well as to my siblings.

TABLE OF CONTENTS

DESCRIPTIONS	PAGE NO.
PERMISSION TO USE	i
ABSTRAK (BAHASA MELAYU)	ii
ABSTRACT (ENGLISH)	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENT	vi
LIST OF FIGURES	viii
LIST OF TABLES	x

CHAPTER 1 : INTRODUCTION

1.1	Background	1
1.2	Problem Statement	5
1.3	Objectives	6
1.4	Scope of the study	7
1.5	Significance of the study	7
1.6	Organization of the report	8

CHAPTER 2 : LITERATURE REVIEWS

2.1	Academic Achievement	9
2.2	Attribute for Measuring Academic Achievement	10
2.3	Data Mining	12
2.4	Data Mining Technique	15
2.5	Conclusion	20

CHAPTER 3 : METHODOLOGY

3.1	Overview of Methodology	22
3.2	Phase 1 - Survey	23
3.3	Phase 2 – CRISP- Data Mining	25
3.4	Phase 3 – Information System Research Design	34
3.5	Conclusion	37

CHAPTER 4 : RESULT AND FINDINGS

4.1	The Sampling Data Set	38
-----	-----------------------	----

4.2	Descriptive Analysis	39
4.3	Predictive Analysis	58

CHAPTER 5 : CONCLUSION

5.1	Conclusion	65
5.2	Future Enhancement	66

REFERENCES	68
-------------------	----

APPENDIXS

Appendix A: Demographics Table	78
Appendix B: Cross tabulation Table	82
Appendix C: Neural Network Result	85
Appendix D: Correlation Result	88

LIST OF FIGURES

Figure 1	The information required for the study	5
Figure 2	Data Mining Technique and Tasks	15
Figure 3.1	Flows for this study	22
Figure 3.2	Example of questionnaire in section Demographics	23
Figure 3.3	Example of questionnaire in section Curriculum	24
Figure 3.4	Phase of the CRISP-DM Reference Model	25
Figure 3.5	Example of questionnaire in part curriculum	27
Figure 3.6	Example of questionnaire in part facilities	27
Figure 3.7	Sample of raw data	39
Figure 3.8	The interface of SAS	32
Figure 3.9	Example of data partition	32
Figure 3.10	Selection of logistic regression	33
Figure 3.11	Selection of correlation Bivariate	33
Figure 3.12	Information System Research Design	34
Figure 3.13	Feed Forward Algorithm	36
Figure 4.1	Graph respondent by IPTA	40
Figure 4.2	Graph respondent by Gender	40
Figure 4.3	Respondent by Academic Qualification	41
Figure 4.4	Respondent by Field of Study	41
Figure 4.5	Respondent by CGPA	42
Figure 4.6	Respondent by Mode of Study	42
Figure 4.7	Respondent by Sponsorship	43
Figure 4.8	Respondent by family income	43
Figure 4.9	Number of respondent with level of satisfying for Curriculum	45
Figure 4.10	Distribution of respondents who are extremely satisfied with the curriculum.	46
Figure 4.11	Number of Respondent by CGPA	53
Figure 4.12	Academic Performance by Gender	54
Figure 4.13	Academic Performance by Age	55
Figure 4.14	Academic Performance by Academic Qualification	55
Figure 4.15	Academic Performance by Field of Study	56
Figure 4.16	Academic Performance by Mode of Study	56
Figure 4.17	Academic Performance by Family Income	57
Figure 4.18	The comparison accuracy between NN and Logistic Regression	61
Figure 4.19	Interface of Prediction System	62

Figure 4.20	User key-in the level of satisfied.	63
Figure 4.21	The output that user key-in.	63
Figure 4.22	Result of Prediction System.	64

LIST OF TABLES

Table 3.1	Range of CGPA	24
Table 3.2	List of Variable for Demographic	39
Table 3.3	List of Variable for Curriculum Section	30
Table 3.4	List of Variable for Facilities Section	30
Table 4.1	The related attribute that involved in this study	39
Table 4.2	Number of respondent with level of satisfying for Curriculum	44
Table 4.3	Distribution of respondents who are extremely satisfied with the curriculum.	46
Table 4.4	Distribution of respondents who are not at all satisfied with the curriculum.	47
Table 4.5	Number of respondent with extremely satisfied level on curriculum items by CGPA	48
Table 4.6	Number of respondent with not at all satisfied level on curriculum items by CGPA	49
Table 4.7	Mode of Study by level of satisfaction with Curriculum	50
Table 4.8	Number of respondent with level satisfying for Facilities	51
Table 4.9	The significant Value for Each Attribute	58
Table 4.10	Academic achievement for NN Model	59
Table 4.11	Academic Achievement for Logistic Regression Model	61
Table 4.12	Comparative accuracy between Neural Network and Logistic Regression	61

CHAPTER ONE

INTRODUCTION

This chapter presents the overview of the study that involves the curriculum and facilities are effect the academic achievement is applying the data mining technique. In addition, this chapter also includes the problem statement, objectives, scope, and significance of study and thesis organization.

1.1 BACKGROUND

Data mining is defined as the extraction from large amount of data. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouse (Berson, 2000). According to Tsantis and Castellani (2001), data mining applies modern statistical and computational technologies to the problem of finding useful patterns hidden within large database. In addition, data mining uses a combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge to uncover hidden trends and

The contents of
the thesis is for
internal user
only

References

- Adewale, O.S., Adebisi, A.B., & Solanke, O.O. (2007) Web-Based Neural Network Model for University Undergraduate Admission Selection and Placement. *Journal of Science and Technology*, 8, 367-384
- Akey. T. M. (2006) *School Context, Student Attitudes and Behaviours, and Academic Achievement*, An Exploratory Analysis , MDRC.
- Averett. S. L., & McLennan, M. C. (2002). Exploring the effect of class size on student achievement: What have we learned over the past two decades?, *Department of Economic and Business Lafayette College Easton*.
- Barker, K., Trafalis, T. & Rhoads, T. R. (2004). Learning From Student Data. *Proceedings of the 2004 System and Information Engineering Design Symposium*.
- Beikzadeh, M. R., & Delavari, N. (2004) A New Analysis Model for Data Mining Processes in Higher Educational Systems. *Faculty of Information Technology, MMU, Cyberjaya, Malaysia*.
- Berson, A., Smith, S., & Therling, K. (2000). An Overview of Data Mining at Dun and Bradstreet. Retrieved from <http://www.theartling.com/text/wp9501/wp9501.htm>.

- Brookshire, R. G., Crews, T. B., Brown, H. F. (2006) Factors Contributing to Student Success in a University Introductory Networks and Telecommunications Course, *Technology Support and Training Management Program*.
- Brown, J. D. (2007) Neural Network Prediction of Math and Reading Proficiency as Reported in the Educational Longitudinal Study 2002 Based on Non- Curricular Variables, Ph.D Dissertation, Duquesne University.
- Chamillard, A.T. (2006) Using Student Performance Predictions in a Computer Science Curriculum. *ITiCSE'06, June 26–28, 2006, Bologna, Italy*.
- Chang, H.C. and Hsu, C.C. (2005) Using Topic Keyword Clusters for Automatic Document Clustering. *Proceedings of the Third International Conference on Information Technology and Applications (ICITA '05)*.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). CRISP-DM 1.0 Step-by-step Data Mining Guide. Retrieve from <http://www.crisp-dm.org/CRISPWP-0800.pdf>.
- Chee, K. H. (2005) Gender differences in the academic ethic and academic achievement , *Project Innovation (Alabama)*.
- Choi, Y (2006) Academic Achievement and Problem Behaviours among Asian

Pacific Islander American Adolescents. *School of Social Service Administration, University of Chicag.*

Cripps.A (1996). Using Artificial Neural Nets to Predict Academic Performance. *Proceedings of the 1996 ACM symposium on Applied Computing, Pp;33-37.*

Duncan, G. J., Dowsett, C. J., Claessens, Amy., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Gunn, J. B., Sexton, H. & Duckworth, K. (2007). School Readiness and Later Achievement. *Developmental Psychology Copyright 2007 by the American Psychological Association, Vol. 43, No. 6, 1428–1446.*

Dunham, H. M. (2003). Data Mining Introductory and Advanced Topics. Introduction.pp. 1-19

Earthman, G. I. (2002) School Facility Conditions and Student Academic Achievement, *Virginia Polytechnic Institute.*

Fan, Y., Zheng, Q. & Xiao, L. J. (2003) Data Mining Application Issues in Fraudlent Tax Declaration Detection. *Proceedings of the Second International Conference on Machine Learning and Cybernetics*, pp. 2202-2206, Xi'an.

- Finn, J. D., Gerber, S. B., Zaharias, J. B. (2005) Small Classes in the Early Grades, Academic Achievement, and Graduating From High School, *Journal of Educational Psychology* 2005, Vol. 97, No. 2, 214–223.
- Gasar, S., Bohanec, M. & Rajkovic, V. (2002) Combined Data Mining and Decision Support Approach to the Prediction of Academic Achievement. Workshop on Integrating Aspects of Data Mining, Decision Support and Meta Learning, Pp; 41-52.
- Ghani, R. & Soares, C. (2006) Data Mining for Business Application. KDD-2006 Workshop, vol. 8, pp. 79-81.
- Golding, P. & Donaldson, O. (2006) Predicting Academic Performance. Proceeding on the 36th *ASEE/IEEE Frontiers in Education Conference, October 28 –31*.
- Golding, P. & McNamarah, S. (2005) Predicting Academic Performance in the School of Computing & Information Technology (SCIT). *ASEE/IEEE Frontiers in Education Conference 2007*, pp. S2H16-T2G24.
- Halawah, I. (2006) The effect of motivation, family environment, and student characteristics on academic achievement, *George Uhlig Publisher*.

Huesman, R. L., Jr., Brown, A. K., Lee G. (2007) Modeling Student Academic Success: Does Usage of Campus Recreation Facilities Make a Difference?, *Paper presented at The National Symposium on Student Retention, Milwaukee September 24-26, 2007.*

Hussain, T. (2007) Student Achievement In Saudi Arabia: The Importance Of Teacher Factors. *Faculty of the Graduate School of Arts & Sciences at Georgetown University.*

Jain, A.K., Duin, R.P.W., & Mao, J. (2000). Statistical Pattern Recognition: A review. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 22 (1): 4-37.

Kusiak, A., Kernstine, K.H., Kern, J.A., McLaughlin, K.A., & Tseng, T.L. (2000).Data Mining: Medical and Engineering Case Studies. *Proceedings of the Industrial Engineering Research 2000 Conference, Cleveland, Ohio, May21-23, 2000, pp. 1-7.*

Lang, C., McKay, J. & Lewis, S. (2007). Seven Factors that Influence ICT Student Achievement, *ITiCSE'07, June 23-27, 2007, Dundee, Scotland, United Kingdom.*

Lonsdale, M. (2003) Impact of School Libraries on Student Achievement. Report for the Australian School Library Association.

- Lipscomb, S. (2005) *Secondary school extracurricular involvement and academic achievement: A fixed effects approach*, Department of Economics, University of California
- Luan, J. (2001). Data Mining As Driven by Knowledge Management in Higher Education. *SPSS Public Conference, UCSP*.
- Luan, J. (2002). Data Mining Application in Higher Education. *SPSS Executive Report*. Retrieve from <http://www.crisp-dm.org/CRISPWP-0800.pdf>.
- Luan, J. (2004). Data Mining and Knowledge Management in higher Education Potential Application. *Proceedings of Air Forum, Toronto, Canada*.
- McMullen, S. (2007) The Impact of Homework Time on Academic Achievement. *The University of North Carolina at Chapel Hill*.
- Mitkas, P. A., Symeonidis, A.L., Kehagias, D. & Athanasiadis, I. (2003) Application of Data Mining and Intelligent Agent Technologies to Concurrent Engineering. *Informatics and Telematics Institute, Greece*.
- Minaei-Bidgoli, B., Kashy, A. D, Gerd Kortemeye, Punch, W. F. (2003)

Predicting Student Performance : An Application of Data Mining Method with an Educational Web-Based System., *Frontiers in Education Conference*.

Naik, B. and Ragothaman, S. (2004). Using Neural Networks to Predict MBA Student Success. Retrieve from
http://findarticles.com/p/articles.ml_m0FCR is_1_38_d_06073208_pg_1.

Nathan, Janet , Robins, A., & Hannah, R. (2004) Interacting Factors that Predict Success and Failure in a CS1 Course, *The SIGCSE Bulletin*, Volume 36, Number 4.

Nguyen, T. N., Janecek, P. & Haddawy, P. (2007). A Comparative Analysis of Technique for Predicting Academic Performance. *ASEE/IEEE Frontiers in Education Conference 2007*, pp. T2G7-T2G12.

Ogor, E. N. (2007) Student Academic Performance Monitoring and Evaluation Using Data Mining *Technique. IEEE Computer Society, pp. 354-359*.

Olson, S. L., Kellum, S. (2003) The Impact of Sustainable Buildings on Educational Achievements in K-12 Schools, *Leonardo Academy Cleaner and Greener Program Report*.

Young, E., Green, H. A., Patrick, L. R., Gibson, T (2003)
Do K-12 School Facilities Affect Education Outcomes?, *A Staff Information Report*.

Romanowski, C. J. (2001) A Data Mining for Design and Manufacturing:

methods and Application. *ACM*, pp. 161-178.

Remondino, M. & Correndo, C. (2006) Mabs Validation through Repeated Execution and Data Mining Analysis. *Journal of Simulation*, 7, 10-21

Roiger, R.J. & Geatz, M. W. (2003). *Data Mining A Tutorial-Based Primer*, 1st edition, *Pearson Education, Inc.*

Rumberger, R. W. (2002) Student Mobility and Academic Achievement, *Clearinghouse on Elementary and Early Childhood Education*.

Salazar, A., Gosalbez, J., Bosch, I., Miralles, R. & Vergara, L. (2004) A Case Study of Knowledge discovery on Academic Achievement, Student Desertion and Student Retention. *Departamento de Comunicaciones Universidad Politecnica de Valencia Spain*.

Siraj, F, Yusof, N. (2006) Web-Based Academic Program Forecasting Model for SPM Certificate Holders Using Neural Network, University Grant Research.

Seifert, J. W. (2004). Data Mining: An Overview. *CRS Report for Congress*.

Sheel, S. J., Renner, R.S. & Dawsey, S. K. (2001). Alternatives to Math Placement Exams: A Look at Discriminant Analysis, Neural Networks, and Ensembles of Networks. *Dept. of Computer Science Dept. of Computer Science AVX Corporation*.

Sepian, N.W. (1981). Conception Educational Achievement.

Educational Reseacher: Vol. 19, No 3.

- Shin, H. M. & Jeong, D. H. (2001) An Application of Data Mining for Marketing in Telecommunication. *Portland International Conference*, vol.1, pp.247-253.
- Superby, J.F. and Vandamme, J.P. (2005) Determination of factors influencing the achievement of the first- year university students using data mining methods. *Production and Operations Management Department, Catholic University of Mons*.
- Tastle, W.J., Russell, J. & Wierman, M.J. (2005) A New Measure to Analyze Student Performance Using the Likert Scale. *Proceeding ISECON 2005*, vol.22, pp. 1-7.
- Thuraisingham, B.M. & Geruti, M. C. (2000) Understanding Data Mining and Applying it to Command, Control, Communication and Intelligence Environment. *IEEE International Symposium*.
- Tsantis, L. & Castellani, J. (2001) Enhancing Learning Environments Through Solution-based Knowledge Discovery Tools: Forecasting for Self-Perpetuating Systematic Reform. *JSET E Journal*, 16.
- Two Crows Corporation (2005), Introduction to Data Mining and Knowledge Discovery. <http://www.twocrows.com/intro-dm.pdf>

Uzoka, F. M. & Seleka, G. G. (2005) B2C E-Commerce Development in

Africa: Case Study of Botswana. ACM June, pp. 290-295.

Vaishnavi, V. & Kuechler, W., (2005). Design Research in Information

Systems. Retrieve June, 28, 2007, From,

http://www.isworld.org/Research/design_driskworld.htm.

Wallman, B.B (1973). Dictionary of Behavioral Science. New York: Van

Nostrand Reintold Company.

Zhang, D. & Zhou, L. (2004) Discovering Golden Nuggets: Data Mining in

Financial Application. IEE Transaction on System, Man and

Cybernetics. vol. 34, pp. 513-522.

Zhihua, X. (1998) Statistics and Data Mining. *Department of Information*

System and Computer Science, National University of Singapore.