The Development of a Prediction System for Student Learning Progress Based on Artificial Neural Networks (A Case Study on Study Program of Mathematics and Statistics – The Faculty of Mathematics and Natural Sciences, Universitas Terbuka)

Dwi Astuti Aprijani, Unggul Utan Sufandi
Computer Centre Universitas Terbuka
Jalan Cabe Raya, Ciputat 15418
Indonesia
Tel. +62 21 7490941, Extension 1407. Fax. +62 21 7429749
dwias@mail.ut.ac.id, unggul@mail.ut.ac.id

ABSTRACT

When new students enroll at the university, they need to fill application forms that incorporate any information about themselves such as academic background, permanent mailing address, gender, date of birth, occupation, marital status, etc. However, this information is not utilized well enough by the university to help in overcoming low graduation rates. This research applies Artificial Neural Network (ANN) Multilayer Perceptron to predict progress learning of students using several parameters such as individual parameter (age, gender), environment parameter (marital status, occupation), and academic parameter (entry semester --odd or even--, grade point average in the first semester, the number of credit hours in the first semester, a cumulative grade point average, the total number of semesters completed at the university, and the study program) at Universitas Terbuka (The Indonesian Open University).

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This research also applies sensitivity analysis method to measure the influence of individual input parameter on any one of possible outcomes.

The experiment had been conducted using student data collection from two study program (Mathematics and Statistic). The data was collected from 3517 students, with 81 students finished their degrees. The experiments used 50% of data as training set, 25% as validation set and 25% testing set. Our experiments and simulation results indicated that the sensitivity analysis method was a potential tool to reduce the complexity of ANN Multilayer Perceptron and to increase the generalization. Generalization is the recognition level of neural network toward the given pattern. The results showed that the generalization of the prototype had an accuracy of 0.992 in predicting the correct outcomes of student graduation.

Keywords: *Continuity of Study, ANN Multilayer Perceptron, Early Stopping, Sensitivity Analysis, Data Mining, Categorical Data*

INTRODUCTION

Universitas Terbuka (UT) is an open university, established in 1984 as the 45th state university to provide opportunity and access to university education for in-service teachers, working adults and recent high school graduates. UT was founded as a part of the government’s national strategies to improve participation in higher education. UT system was considered likely to be economical by the Government and accessible by the students. UT has been founded to be complementary to the existing higher education system. UT’s target audience are those opt for distance learning because they have work commitment, reside in locations where there are no access to conventional universities, and they need to have flexibility and freedom from the strict schedules required in classroom-based learning. Therefore there is no admission test to study in UT. The only requirement is that candidate students are graduated from high school. This lead to the much dispersed student abilities.

Research has shown that students in the distance learning (DL) system have low achievement. This problem arises from the inadequacy of students to adapt with the learning model from the guided study to independent study, the communication model from direct face-to-face to long distance one, and conveying model from oral/verbal to written one (by applying technologies), and learning environment from campus-based study to home-based study (Kadarko 2000).

UT statistics at 2005 indicated that from total of 4,013,804 students, more or less 36% (1,458,401) students did not re-register at the next registration session. Especially for Mathematics undergraduate and Statistics undergraduate study programs, those that did not re-register at 2nd semester were 48.3% and 42.8% respectively, and at 3rd semester were 28.9% and 28.5%, respectively.

In regards to the implementation of quality assurance system in UT, it was necessary to determine those factors that affect the learning continuity of students at the next semester. The continuity of student learning progress in this research is categorized into 1) finished (graduated) and 2) unfinished (non-active). The method used in predicting the outcome of the continuity is Artificial Neural Network.

OBJECTIVES

Objectives of this research are to determine factors/variables that affect the learning progress of PTJJ students, especially Mathematics and Statistics students.

Those affecting factors then could be utilized to support the management in applying the quality assurance system for monitoring the students' progress, individually or as whole.

LEARNING PROGRESS

The UT education system does not recognize drop-outs. Students may elect to have academic vacation for 4 registration session in consecutive without reporting to UT, however before the vacation time ends, they have to re-register to maintain their active status as student. Otherwise, their status becomes non-active student. The term of learning progress in this research is identical with the student graduation.

The tendency of decreasing persistency of students was affected by 3 dependent factors: individual, environment, and academic (Belawati, 1997). Other researches showed that the dynamics of learning volition is related to motivation aspects (direction, energizing, and persistence). The high learning progress leads to survival behavior and showed up as student persistence to keep following lectures in the DL institutes (Darmayanti 2002).

Factors of individual/environment, academic services and administrative service would also affect student resistance (Isfarudi, 1994). While Nuraini (1991) concluded that the continuity of UT student registration was related to their previous exam marks and their study program. Students with high exam marks tend to have high registration continuity. According to Zu (2000), satisfying results during the first year of the learning process will lead to positive effects on the learning persistence or study continuity.

ARTIFICIAL NEURAL NETWORK

Artificial Neural Network (ANN) is one of the artificial representations of human brain in order to try to simulate the learning process of the brain. The term artificial is used since this neural network is implemented as computer program that can solve several calculation processes during learning process. ANN is trained to create a reference model and the trained ANN can then be used to recognize patterns (Kusumadewi 2004).

A type of ANN that has showed satisfactory performance is back propagation Multi Layer Perceptron (MLP) with supervised training. According to Kantardzic (2003), MLP has 3 characteristics, which are (1) model of each neuron usually contains non-linear activation function, such as sigmoid or hyperbolic; (2) the network contains one or more hidden layer which are not part of the input nor the output layer; and (3) the network has connections from one layer to another layer.

METHOD AND IMPLEMENTATION

1. **Data Sets**

Data used for training process and model testing is data with category finished or not finished with 1759 items, while prototype testing data used is 879 items (Table 1).

Table 1. Data Category

Category	Data			
	Training Data (50%)	Validation Data (25%)	Testing Data (25%)	Amount
Finished	41	20	20	81
Not finished	1718	859	859	3436
	1759	879	879	3517

2. Research Flowchart

The steps of this research are simplified in this flowchart below.

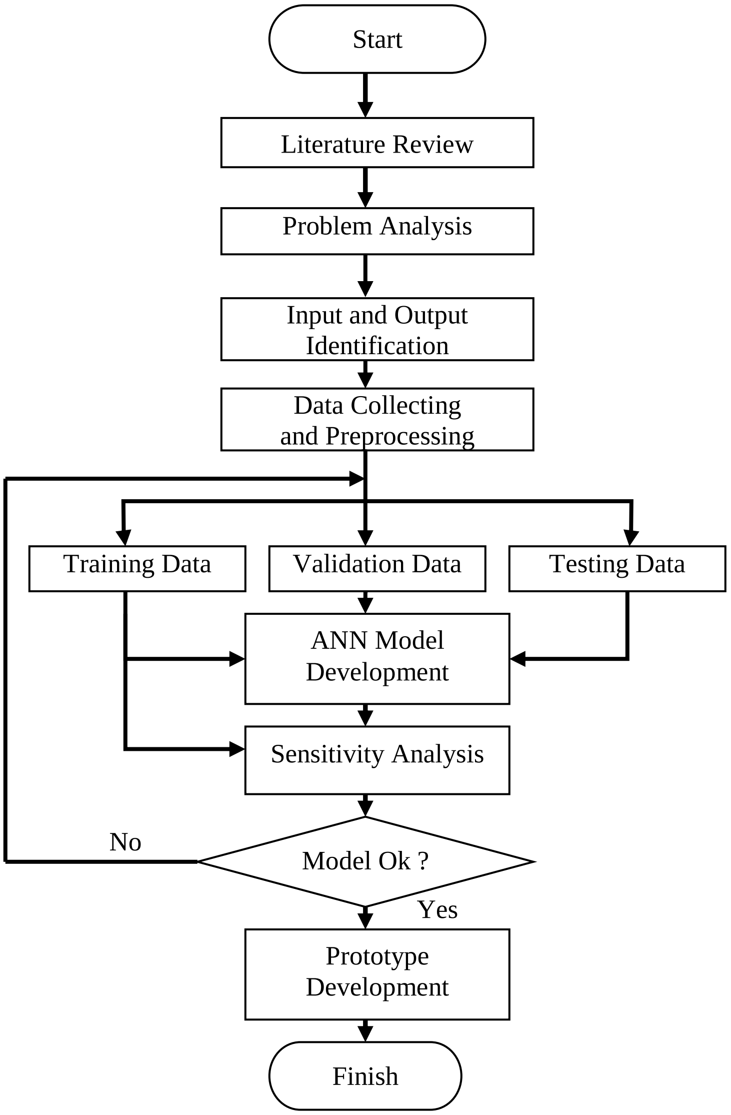

Figure 1. The flowchart of system model development

3. Model Architecture

The system prototype was developed using Matlab version 6.1 for the modeling of ANN, testing and also sensitivity analysis. The user interface was built using Sybase Power Builder version 7.0. This system comprised of 4 modules: training module, ANN analysis (testing) module, sensitivity analysis module and prediction module (Figure 2).

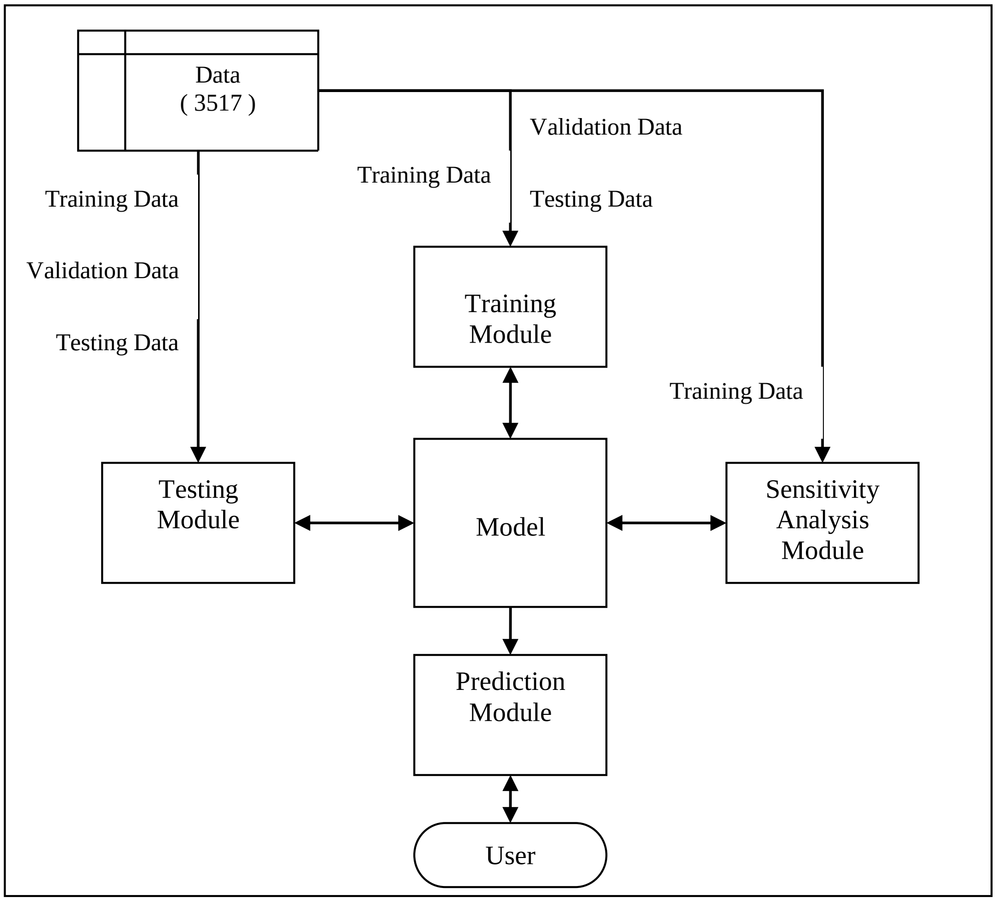

Figure 2. System Model Design

RESULT AND DISCUSSION

Results from pre-processing stage were normalized, unary-encoding data with 11 input variables encoded into 16 input nodes and 2 output nodes. The architecture of ANN used to develop model at 1st iteration was shown in Figure 3.

1. Model Development

The first step is developing ANN with structure as shown in Figure 3. At 1st iteration, all 11 input variables were used which corresponded to 16 input nodes. The behavior of the ANN was then observed by decreasing the number of variables or input nodes using sensitivity analysis.

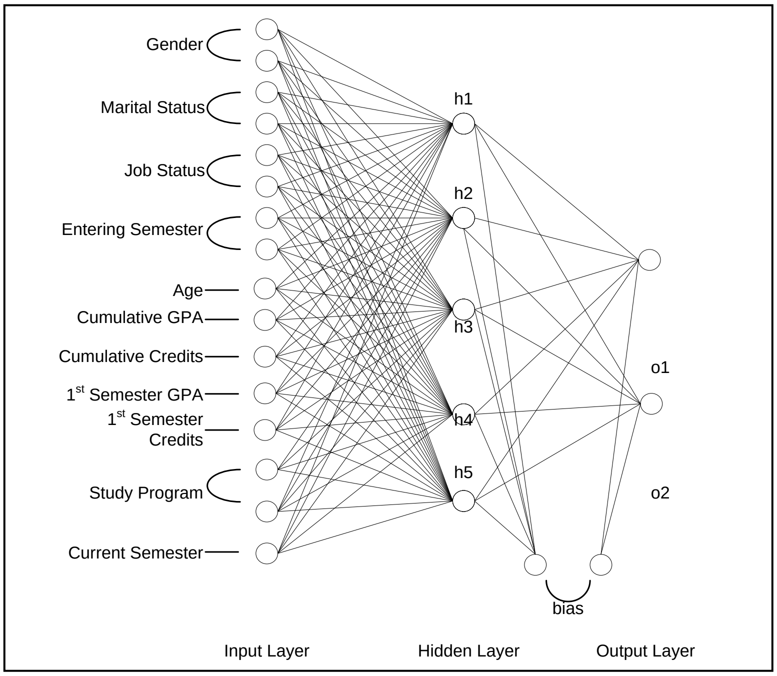

Figure 3. ANN architecture at 1st iteration

The analysis of the performance of the trained ANN toward the testing data was shown in Table 2.

Table 2. Training and testing result at 1st iteration

NI	Training (Training Data)		Testing					
			Validation Data			Testing Data		
	Time (Second)	Epoch	Generalization			Generalization		
			Amount		(%)	Amount		(%)
16	0,9840	67	856 0	0 16	99,20	854 0	0 15	98,86

Table 2 showed that training process at 1st iteration spent 0.9840 seconds. Epoch of 67 means that the validation process using validation data terminates the training process after 67 iterations (epoch) as after the last iteration, the level of generalization to the validation data is considered constant and the validation error starts to increase.

The next step was to perform sensitivity analysis process for the ANN with the best performance where the result of the process can be seen in the following table:

Table 3. Sensitivity analysis result at 1st iteration

No.	Variable	Sensitivity	
		Value	Rank
1.	Gender	0,677642	6
2.	Marital Status	0,823301	7
3.	Job Status	0,823301	7
4.	Entering Semester	0,584603	5
5.	Study Program	0,404459	3
6.	Age	0,41211	4
7.	Cumulative GPA	4,6158	9
8.	Cumulative Credits	5,3418	10
9.	1st Semester GPA	0,11074	1
10.	1st Semester Credits	1,3897	8
11.	Current Semester	0,23765	2

The above table indicated that the variable with the highest sensitivity level was 'cumulative credits' with value of 5.3418, while the lowest was '1st semester GPA' with value of 0.11074. For the next iteration, variables with lowest level of sensitivity, which were '1st semester GPA' and 'current semester' were removed from training, validation and testing data set.

At 2nd iteration, the ANN structure model was derived from the 1st iteration model, but without the input variables of '1st semester GPA' and 'current semester' so the total of input variables were 9 or equivalent to 14 input nodes. The generalization of the validation data and testing data was 99.54% and 99.20%, respectively. The sensitivity analysis and input variable elimination process from 1st iteration increased the generalization of validation and testing data as much as 0.3427% and 0.3439%, respectively.

Result of sensitivity analysis from the 2nd iteration showed that the variable with highest sensitivity level was 'cumulative credits' (6.380733). Variables with lowest sensitivity level are 'job status' (0.131000) and 'age' (0.418034). For the next iteration, these low sensitivity variables were removed from training, validation and testing data set.

The ANN structure used for 3rd iteration was similar as Figure 3 but without the '1st semester GPA', 'current semester', 'job status' and 'age' variables. This ANN had 7 input variables with 11 input nodes. Generalization against validation data was 99.32% while generalization against testing data stayed at 99.20%. The sensitivity analysis and input variable elimination process at 2nd iteration instead decreased the generalization against validation data.

The sensitivity analysis process of the 3rd iteration showed that variable with the highest level of sensitivity was 'cumulative credits' (6.380733). Variables with lowest sensitivity level were 'study program' (0.027609) and 'entering semester' (0.042065). For the next iteration, both lowest sensitivity variables were removed from training, validation and testing data set.

The ANN structure for the 4th iteration was similar as Figure 3, but without '1st semester GPA', 'current semester', 'job status', 'age', 'entering semester' and 'study program' variables, so the total of input variables was 5 or equivalent to 7 input nodes.

The generalization against validation and testing data for this iteration were 97.84% and 98.52%, respectively. The sensitivity analysis and input variable elimination process from the 3rd iteration was, on the contrary, decreased the generalization against validation and testing data even further. Hence, both 'entering semester' and 'study program' variables were reconsidered for re-inclusion into the network.

Table 4. Training and testing result at 4th iteration

NI	Training		Testing					
	(Training Data)		Validation Data			Testing Data		
	Time (Second)	Epoch	Generalization			Generalization		
			Amount		(%)	Amount		(%)
7	1,2970	104	856	0	97,84	858	0	98,52
			0	4		0	8	

At the 5th iteration, the ANN structure was based from the 3rd iteration but without the 'study program' variable so the number of input variables was 6 and input nodes was 9.

Table 5. Training and testing result at 5th iteration

NI	Training		Testing					
	(Training Data)		Validation Data			Testing Data		
	Time (Second)	Epoch	Generalization			Generalization		
			Amount		(%)	Amount		(%)
9	0,8900	61	856	0	99,09	856	0	99,20
			0	15		0	16	

Table 5 showed that generalization of validation data and testing data was 99.09% and 99.20%, respectively. The sensitivity analysis and input variables elimination process at this 5th iteration increased the generalization of validation data and testing data as much as 1.2776% and 0.6902%, respectively. In other words, the inclusion of 'entering semester' variable increased the generalization which showed that it was an influential variable.

Generalization of testing data at 3rd iteration was 99.20%, at 4th iteration decreased to 98.52% and at 5th iteration increased back to 99.20%. Hence, it could be concluded that the maximum generalization of testing data was 99.20%. At 5th iteration, the sensitivity of all variables was high enough and the iteration process was terminated.

Table 6. Sensitivity analysis result at 5th iteration

No.	**Variable**	**Sensitivity**	
		Value	**Rank**
1.	Gender	0,783686	2
2.	Marital status	0,783686	2
3.	Entering Semester	0,042065	1
4.	Cumulative GPA	5,459138	5
5.	Cumulative Credits	6,380733	6
6.	1st Semester Credits	1,462324	4

Hence the variables that affect learning progress of UT student on the Mathematics and Statistics Study Program are gender, marital status, entering semester, cumulative GPA, cumulative credits and 1st semester credits.

2. The Best ANN Architecture

The best ANN architecture from this research shown in Figure 4 was the result of the 5th iteration with 6 input variables and 9 input nodes.

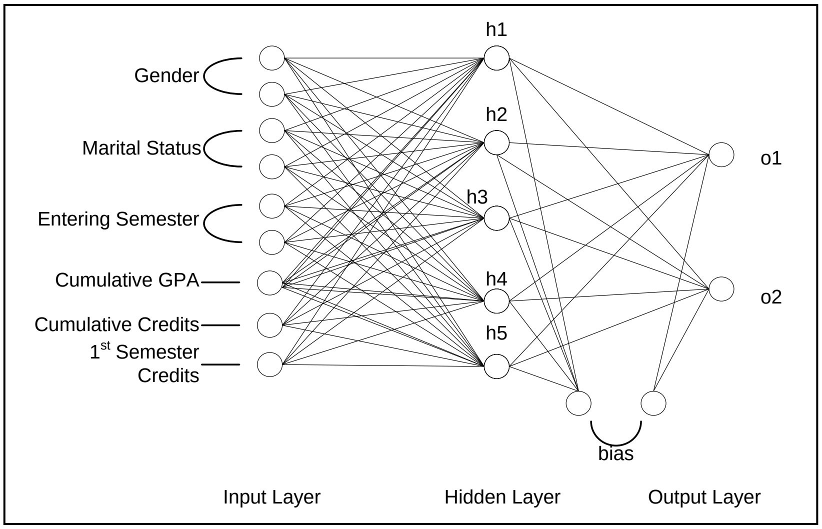

Figure 4. The best ANN architecture

CONCLUSION

- Back-propagation Artificial Neural Network (ANN) is suitable for prediction purposes.
- ANN architecture with the best generalization in this research used composition of 50% training data, 25% validation data and 25% testing data, with 11 input nodes, 5 hidden nodes and 2 output nodes with training rate of 0.001, and had the highest generalization level against testing data and validation data of 99.20% and 99.09%, respectively.
- Sensitivity analysis could decrease the input nodes/variables in order to decrease the complexity of ANN. For this research, the number of variables/input nodes could be decreased from 16 input nodes to 9 input nodes.
- The division of the data set into 3 partitions (training, validation and testing data set) is generally very helpful in controlling the training process of ANN.
- Longest time taken for the training process of the ANN with 3517 input data was 1.2970 second with 104 iterations.

SUGGESTION

- This research may be developed further to obtain a better system.
- Increasing the number of input variables to the model in order to improve the level of accuracy and the generalization.
- Sensitivity analysis emphasizes on the input nodes which can be combined with the other analysis that emphasizes on the hidden layer.
- Sensitivity analysis can be implemented in the trained ANN model successfully, hence further research with different composition of training, validation and testing data (such as 60%-20%-20% or 70%-15%-15%) and with 5 nodes in the hidden layer and different rate of training can be performed.

REFERENCES

Belawati, T. 1997, Understanding and Increasing Student Persistence in Distance Education: A Case of Indonesia, *JURNAL STUDI INDONESIA* 1997:7:1 [terhubung berkala] http://pk.ut.ac.id/jsi/ 71tian.htm [27 Maret 2006]

Darmayanti, T. 2006. Kemauan Belajar (Learning Volition) Mahasiswa Pendidikan Jarak Jauh (Studi Kasus di Universitas Terbuka), *JURNAL PENDIDIKAN TERBUKA DAN JARAK JAUH* 2002:3:1 [terhubung berkala] http://pk.ut.ac.id/ptjj/31darmayanti.HTM [28 Maret 2006]

Engelbrecht,AP. Cloete, I. Zurada, JM. 1995. *Determining the Significance of Input Parameters using Sensitivity Analysis.* College of Information Sciences and Technology. [terhubung berkala]. http://citeseer.ist.psu.edu/ rd/22639223%2C464485%2C1%2C0.25%2CDownload/http://citeseer.ist.psu.edu/ cache/papers/cs/25171/http:zSzzSzwww.cs.up.ac.zazSz%7EengelzSzpublicationsz SzIWANN95a.pdf/engelbrecht95determining.pdf [06 Juni 2006]

Han, J. Kamber, M. 2001. *Data Mining: Concept, Model, Methods, and Algorithm*. Wiley-Interscience, New Jersey.

Isfarudi, 1994. *Faktor-faktor Penentu Resistensi Belajar Mahasiswa FMIPA Universitas Terbuka*. Tesis. IKIP Jakarta, Jakarta.

Kantardzic, M. 2003. *Data Mining: Concept and Techniques*. Morgan Kaufmann Publisher, San Fransisco.

Kusumadewi, S. 2004. *Membangun Jaringan Saraf Tiruan (Menggunakan Matlab dan Excel Link).* Graha Ilmu, Yogyakarta.

Mathworks, Inc. 2001. *Sample Training Session*: *Matlab Documentation Version 6.1.0.450 Release 12.1.*

Nuraini. 1991. *Kontinuitas Registrasi dan Hubungannya dengan Nilai Ujian yang Diperoleh.* Universitas Terbuka, Jakarta.

Supratman, A, Zuhairi, A. 2004. *Pendidikan Jarak Jauh: Teori dan Praktek*. Pusat Penerbitan Universitas Terbuka, Jakarta.

UT. 2002. Jaminan Kualitas pada Pendidikan Tinggi Jarak Jauh di Indonesia, *JURNAL PENDIDIKAN TERBUKA DAN JARAK JAUH* 2002:3:1 [terhubung berkala] http://pk.ut.ac.id/ptjj/31simintas.HTM [30 Maret 2006]

UT. 2005. *Katalog UT 2005-2006.* Jakarta : Pusat Penerbitan

UT. 2005. *Statistik UT 2005.* Jakarta : Pusat Penerbitan

UT. 2006. *Katalog UT 2006.* Jakarta : Pusat Penerbitan

Yao, J.T. 2003. *Sensiti*vity *Analysis for Data Mining.* Proceeding of 22nd International Conference of North American Fuzzy Information Processing Society - NAFIPS. Chicago. Illinois. 24 – 26 Juli 2003. halaman 420 – 425.

Zu, Lillian. 2000. *How the First-year College Experience Contribute to The Persistence*. SUNY College. Brockport. [terhubung berkala]. http://www.ocair.org/files/Presentations/onlinepapers/LilianZhu.pdf [26 Maret 2006]