Data Preparation in Predictive Learning Analytics (PLA) for Student Dropout

Nurmalitasari, Zalizah Awang Long, Mohammad Faizuddin Mohd Noor

Abstract: ***Predictive learning analytics (PLA) are the current trend to support learning processes. One of the main issues in education particularly in higher education (HE) is high numbers of dropout. There are little evidences being identified the variables contributing toward dropout during study period. The dropout are the major challenges of educational institutions as it concerns in the education cost and policy-making communities. The paper presents a data preparation process for student dropout in Duta Bangsa University. The number of students dropout in Duta Bangsa University are in high alarm for both management and also educator in Duta Bangsa. Preventing educational dropout are the major challenges to Duta Bangsa University. Data preparation is an important step in PLA processes, the main objective is to reduce noise and increase the accuracy and consistency of data before PLA executed. The data preparation on this paper consist of four steps: (1) Data Cleaning, (2) Data Integration, (3) Data Reduction, and (4) Data Transformation. The results of this study are accurate and consistent historical dropout data Duta Bangsa University. Furthermore, this paper highlights open challenges for future research in the area of PLA student dropout.***

Keywords: Data Preparation, Dropout, PLA, Duta Bangsa University.

I. INTRODUCTION

Education era 4.0 is a phenomenon that define again the education basis by placing students at the middle of the ecosystem and changing the focus of teaching to learning. PLA are the current trend to support learning processes. One of the main issues in education particularly in higher education (HE) is high numbers of dropout. Dropping is a important problem in higher education and the policymaking community [1]. There have been a lot of studies on student dropout data analysis i.e, [1][2][3][4][5][6]. However, nothing has been discussed about the detailed data preparation analysis. Whereas data preparation is a fundamental stage of data analysis in PLA for student dropout.

Based on Higher Education Data Base (Pangkalan Data Perguruan Tinggi/ PDDIKTI) data (2017 & 2018), in Indonesia within the last two years, the percentage of the number of students dropout was getting bigger. In 2017, the percentage of students dropout of the total number of students was 2.8%, with the number of student dropouts being 195,176 student. In 2018, the percentage of student dropout of the total number of student was 3%, with the number of student dropouts being 245,810 student. Duta Bangsa University is a private university that developed in Central Java. The number of dropout students at Duta Bangsa University Faculty of Computer Science is very high. The ratio of enrollment students with the number of students dropping out in the academic year from 2011 to 2015 respectively 33.87%, 44.52%, 45.21%, 51.32% and 48.52% which looks up year after year (Academic Department of Duta Bangsa University, 2019). The number of dropout students at Duta Bangsa University are in high alarm for both management and also educator in Duta Bangsa. Therefore it is important to analyze student dropout data to determine the policies that can be taken by Duta Bangsa University.

In this paper present data preparation process for dropout students at Duta Bangsa University. The purpose of doing this data preparation are: (1) To make it easier to understand data so as to facilitate the selection of techniques and PLA methods that are appropriate for analyzing student dropouts; (2) to improve data quality so that the results of PLA process get better; and (3) To increase the efficiency of PLA for student dropout.

II. BACKGROUND THEORY

A. Data Preparation

Data Preparation is a process/step taken to make raw data into quality data (good input for data mining tools). There are four stages in data preparation, i.e.: (1) data cleaning, (2) data integration, (3) data reduction, and (4) data transformation.

- ***Data Cleaning***

At present, the amount of data on the education system is increasing exponentially, fast extraordinary. But in general the data has noise, is incomplete and inconsistent. Data cleaning (or data cleansing) routines attempt to fill up in missing values, Noisy data, and correct incompatibility in the data. ***Missing value,*** there are several ways to clear missing values [7], i.e: (1) disregard the tuple; (2) Fill up in the missing value manually; (3) Use a universal constant to fill up in the missing value; (4) Use a quantify of central tendency for the attribute; (5)

Revised Manuscript Received on January 5, 2020.
***** Correspondence Author
Nurmalitasari*, Malaysian Institute Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia. Email: nurmalitasari@udb.ac.id.
Zalizah Awang Long, Malaysian Institute Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia. Email: zalizah@unikl.edu.my
Mohammad Faizuddin Mohd Noor, Malaysian Institute Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia. Email: mfaizuddin@unikl.edu.my

Use the mean or median of an attribute to fill up all samples in the same type as the tuple; (6) Use the most possible value to fill up in the missing value.

Noisy Data, noisy in a data set can be an error or variance that is random. For example, a value that is much smaller or larger than others. There are several ways to smooth out noisy data i.e: Binning, Regression, and Outlier analysis. In outliers analysis one of the methods used to detect outliers is z-score [8]. The Z-score is computed as

$$Z = \frac{x_i - \bar{x}}{\sigma} \qquad (1)$$

where x_i is value of variables, $\bar{x}$ is average of these variables, and σ is standard deviation of the data.

- ***Data Integration***

One important problem in data integration is redundancy. One strategy to overcome redundancy is to use correlation analysis. Correlation analysis is used to determine the relationship between one variable with other variables based on existing data. The χ^2 (chi-square) is used to analyze the correlation of nominal data [9]. The Pearson's product moment coefficient is used to analyze numeric data [10]. The χ^2 (chi-square) is computed as

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}} \qquad (2)$$

where o_{ij} is the observed frequency, and e_{ij} is the expected frequency. The Pearson's product moment coefficient is computed as

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \qquad (3)$$

Where *n* is number of data and *x*, *y* are the variable to look for correlation.

- ***Data Reduction***

The data reduction stage is used to get representations of smaller data sets, but still maintain the integrity of the original data. Data reduction means reducing the size of the data set but producing the same analysis results. Data reduction is divided into several groups namely dimensional reduction, data reduction and data compression. One of the most popular strategies for reducing data is Principal Component Analysist (PCA). PCA is a technique to cut data dimensions, increase interpretability and minimize information loss. The detailed steps for data reduction using PCA can be seen at Ian [11].

- ***Data Transformation***

Data transformation is a process of transformation or consolidation of data into the required forms of mining. One of the data transformation strategies is normalization. Normalization data is used to change the scale of the data into smaller ranges. Many strategies are used for data normalization, but in this study will use the Z score normalization method accordingly [8] the same as the equation (1).

B. Duta Bangsa University

Duta Bangsa University is a private university that developed in Central Java. Duta Bangsa University was established in 2018. Duta Bangsa University is a combination of Duta Bangsa Collage which joins the Academy of Medical Records and the Midwifery Academy Citra Medika. Duta Bangsa University is ranked 788 in Indonesia. The number of dropout students at Duta Bangsa University Faculty Computer Science is very high. Duta Bangsa University Faculty of Computer Science has 4 study programs, including information systems, informatics engineering, informatics management and computer engineering. There are 4 class programs there namely regular classes, employee classes, week end classes, and distance classes. In the Duta Bangsa University rule, the student is in dropout when: (1) student submits resignation; (2) the student died; (3) more than four semesters without information to the campus; and (4) exceeded study limit (seven years).

III. METHODS

The data used in this study were 406 data student dropouts from Duta Bangsa University Faculty Computer Science from 2014 to 2018. Data obtained from Academic Department of Duta Bangsa University. These data include study programs, gender, grade point academic (GPA), class programs, school origin, age, parent jobs and addres. The stages of data preparation in this study can be seen in full in the following Figure 1.

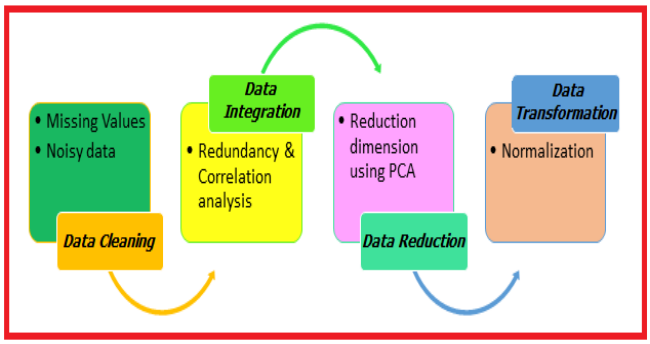

Fig. 1. The stages of data preparation student dropout

IV. RESULT AND DISCUSSION

Student dropout data obtained from the Academic Department of Duta Bangsa University are highly susceptive noisy and missing data, and will produce inferiority data in the results of PLA. Therefore it is very important to conduct the data preparation stage before conducting further research on student dropouts.

A. Data Cleaning

The variables analyzed in this study are study program variables, gender, GPA, class program, school origin, address, parent jobs and age. Variable names and ID are not included in the analysis because they are considered to have no effect on the dropout.

Student data obtained from the UDB academic department a lot of missing value, that is, school origin 9 data missing, address 19 data missing, parent jobs 80 data missing, and age 5 data missing. To overcome this missing value, we use a value from the central tendency, which is the median. The median is chosen because the school origin, address, parent jobs, age data have asymmetrical data.

Dropout student data Duta Bangsa University also contains noise. One of the techniques used to overcome noise is outlier analysis. In this study the outlier analysis using Z-score. The data is said to be outlier if the value of z is greater than z = 2.5 or smaller than z = -2.5 and outlier data discarded. The eight variables used in this study there were only three variables containing outliers, namely study program, GPA and age variables (can be seen in Figure 2, 3, and 4). Z-score value calculation using equation (1). The result of outlier analysis can be seen in Figure 5, 6 and 7.

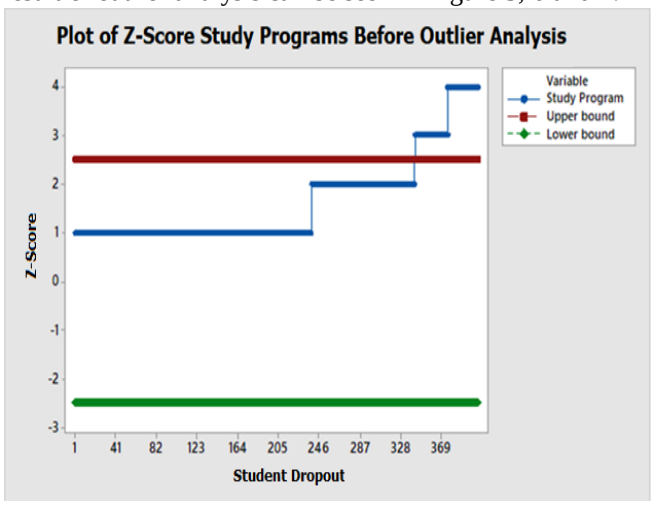

Fig. 2. Plot of Z-Score Study Programs Before Outlier Analysis

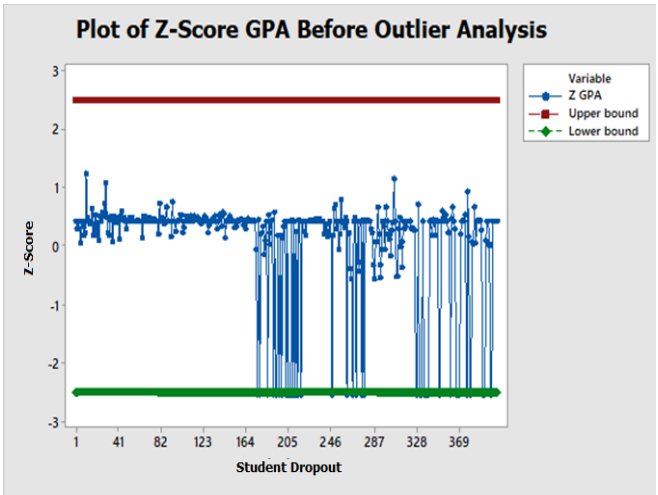

Fig. 3. Plot of Z-Score GPA Before Outlier Analysis

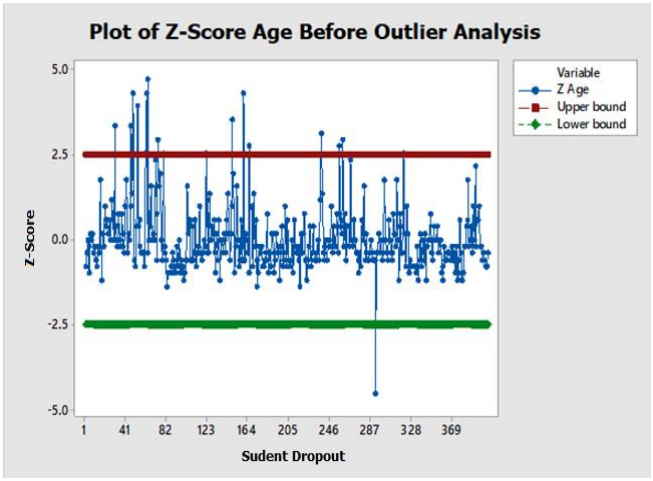

Fig. 4. Plot of Z-Score Age Before Outlier Analysis

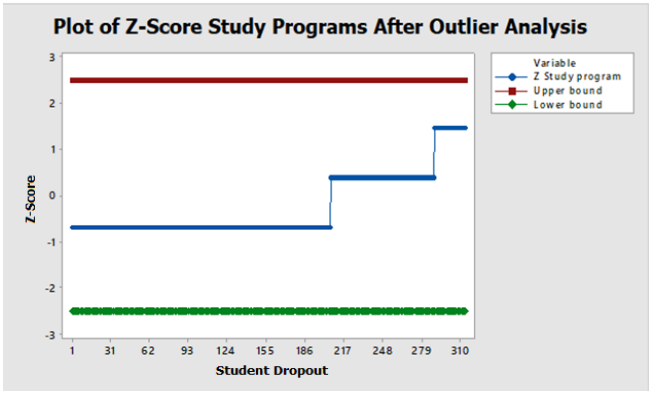

Fig. 5. Plot of Z-Score Study Programs After Outlier Analysis

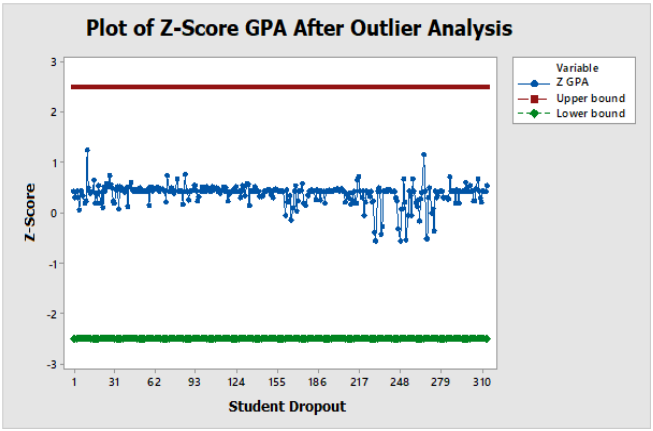

Fig. 6. Plot of Z-Score GPA After Outlier Analysis

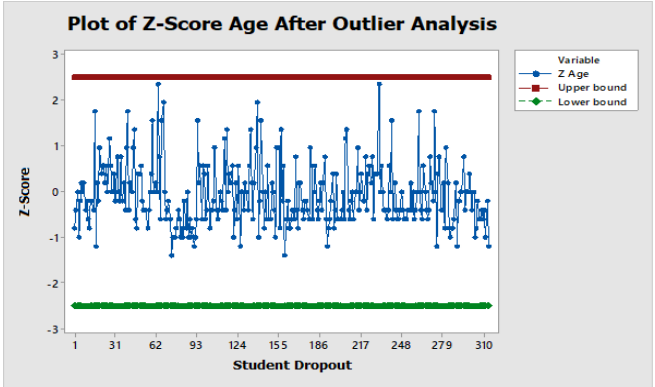

Fig. 7. Plot of Z-Score Age After Outlier Analysis

B. Data Integration

The data integration stage discussed in this research is redundancy. The technique used to detect redundancy are correlation chi-square (using equation (2)) for data nominal, Pearson for data numeric (using equation (3)) and Tau-Kendal for combination data numeric and nominal. The significance value (Sig.) of the correlation test results are presented in the following Table-I. To reduce data redundancy, two correlated variables can be represented by only one variable, in other words the other variable is discarded. Based on Table-I it can be concluded that the variables of study programs, gender, class programs and Age can be eliminated.

Table-I. The significance value (Sig.) of the correlation test results

	Study program	Gender	GPA	Class Program	School Origin	Address	Parent Jobs	Age
Study program		0.000	0.000	0.000	0.000	0.135	0.003	0.727
Gender	0.000		0.023	0.655	0.893	0.385	0.217	0.218
GPA	0.000	0.023		0.121	0.685	0.791	0.270	0.496
Class Program	0.000	0.655	0.121		0.000	0.000	0.003	0.000
School Origin	0.000	0.893	0.685	0.000		0.469	0.079	0.000
Addres	0.135	0.358	0.791	0.000	0.469		0.240	0.514
Parent Jobs	0.003	0.217	0.27	0.003	0.079	0.24		0.273
Age	0.727	0.218	0.496	0.000	0.000	0.514	0.273	

C. Data Reduction

There are four variables analyzed using PCA, among them are GPA, School Origin, Addres, and Parent Jobs. PCA calculation using SPSS software.

Table-II KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.511
Bartlett's Test of Sphericity	**Approx. Chi-Square**	9.132
	df.	6
	Sig.	

The KMO test aims to ascertain whether the Synthetic Aperture Personality Assessment (SAPA) data to be analyzed is considered sufficient, so that it can be used in factor analysis. Based on Table-II, can be seen the value of KMO =0.505 > 0.5, meaning that by using a 95% confidence level it can be concluded that the amount of SAPA data has been sufficiently factored. Bartlett's test is used to determine whether there is a relationship between variables.

The scree plot in Figure 8 is a plot of eigenvalue as a function of the number of factors in an attempt to extract. The form of the scree plot is used to determine the number of factors. The number of factors is indicated by the initial eigenvalue value greater than 1. Based on Figure 8 the initial eigenvalue that is greater than 1 is component number 1and 2, each component is described in the Table-III.

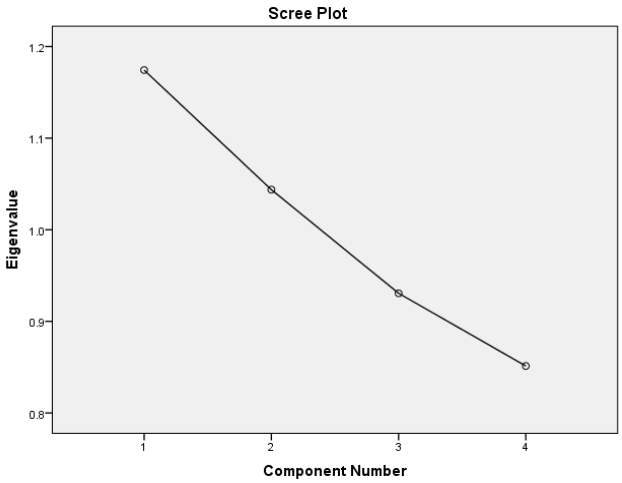

Fig. 8. Scree Plot of Eigen Value

Tabel-III Component Matrixa

	Component	
	1	**2**
Grade Point Average	.496	-.463
Parent Jobs	.723	.159
Address	.575	.540
School Origin	-.274	.716

Extraction Method: Principal Component Analysis
[a] 2-component extracted

Tabel-IV Component Matrixa

	Component	
	1	**2**
Grade Point Average	.185	-.653

Parent Jobs	.701	-.238
Address	.772	.164
School Origin	.136	.754

Extraction Method: Principal Component Analysis
Rotation Method: Varimax with Kaiser Normalization
[a] Rotation converged in 3 iterations extracted

Through the rotation of the variable distribution it becomes increasingly clear the factors obtained. Based on Table-IV after rotation there are two variables that are highly correlated (cut off point = 0.55) with factor 1 consists of parent job and addres, this factor is called Environmental factor. Factor 2 consists of GPA and School Origin, this factor is called Education factor.

D. Data Transformation

The normalization method used in this study is the Z Score. The Z-score calculation using equation (1). The results of normalization data GPA, school origin, address, dan parent jobs variables are presented in the Figure 9.

Data preparation for student dropout has been carried out and obtained higher quality data results because the data does not contain missing values, noise, redundancies, excessive data dimensions and the data are normal.

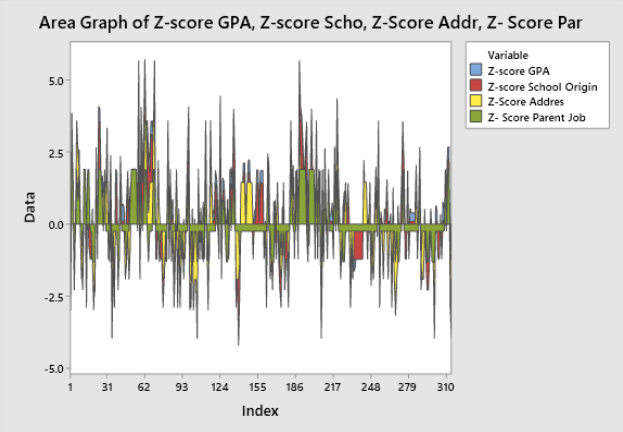

Fig. 9. Z-Score of variable GPA, school origin, Addres, and Parent job

V. CONCLUSION

Data preparation for Analyzing dropout students at Duta Bangsa University has been successfully carried out. The results of this study are accurate and consistent historical dropout data Duta Bangsa University. Furthermore, this paper highlights open challenges for future research in the area of PLA student dropout.

REFERENCES

1. Aulck, L, Velagapudi, N, Blumenstock, J and West, J. 2016. Predicting Student Dropout in Higher Education. *In: ICML Workshop on #Data4Good: Machine Learning in within the Open Polytechnic of New Zealand, relying Social Good Applications*. New York, NY, USA
2. Pedro A. Willging, Scott D. Johnson, 2004. Factors that influence students' decision to drop out of online courses. *Journal of Asynchronous Learning Network*. Vol. 8, No.3, pp.115-127.
3. Gerben W. Dekker, Mykola P., & Jan M. V., 2009. Predicting Students Drop Out: A Case Study. *Educational Data Mining*, pp. 41-50.
4. Allan Sales, Leandro B., & Adalberto C., 2015. Predicting Student Dropout: A Case Study in Brazilian Higher Education. *3rd KDMiLe – Proceedings* – ISSN 2318-1060, Petropolis, RJ, Brazil.
5. Vlatko N., Riste S., Igor M., & Ivan C., 2015. Educational Data Mining: Case Study for Predicting Student Dropout in Higher Education, *12th International Conference on Informatics and Information Technologies Proceedings.* Available: https://www.researchgate.net/publication/282333827
6. Subitha S., Sivakumar V., and Rajalakshmi S. 2016. Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree. *Indian Journal of Science and Technology*. Vol. 9, No. 4. Available: DOI: 10.17485/ijst/2016/v9i4/87032
7. Han, J., Kamber, M., and Pei, J., 2012. Data Mining: Concepts and techniques. Third Edition. Elsevier.
8. Sushree Swarupa T., Rajiv Kumar S., and Prabhat Kumar G. 2013. Comparison of Statistical Methods for Outlier Detection in Proficiency Testing Data on Analysis of Lead in Aqueous Solution. *American Journal of Theoretical and Applied Statistics*. Vol.2, No.6, pp.233-242.
9. Mary L. M. 2013 Juny. The Chi-square Test of Independence. *Biochemia Medica*. Vol.23, No.2, pp.143-149.
10. David A. Walker, 2017. The Pearson Product-Moment Correlation Coefficient and Adjustment Indices: The Fisher Approximate Unbiased Estimator and the Olkin-Pratt Adjustment (SPSS). *Journal of Modern Applied Statistical Methods*. Vol.16. No.2. pp.540-546.
11. Ian T. Jolliffe and Jorge Cadima, 2016. Principal component analysis: a review and recent developments. Phil. Trans. R. Soc. A 374: 20150202. Available: http://dx.doi.org/10.1098/rsta.2015.0202.

AUTHORS PROFILE

Nurmalitasari, is mathematics lecturer at Duta Bangsa University Faculty Computer Science, Indonesia. She is currently a PHD student at the Universiti Kuala Lumpur (UniKL) majoring in information technology. Research Areas: Forecasting, Fuzzy Time Series, Neural Network, Learning Analysis, Statistical Inferences, Time series and Artificial Intelligence.

Assoc. Prof. Dr. Zalizah Awang Long, is currently the Dean of Universiti Kuala Lumpur, Malaysian Institute of Information Technology (MIIT). She holds a PHD in Data Mining in public health from the Universiti Kebangsaan Malaysia, (UKM) Bangi in 2012. She worked for UniKL MIIT for 19 years from June 1999 until present and held some important and key positions. Research Areas: Data Mining, Artificial Intelligent, Database, Software development, Management Information System, E-Commerce, and Public Health Application.

Dr. M. Faizuddin M. Noor, is a Senior Lecturer in the Software Engineering Information System Department at Universiti Kuala Lumpur Malaysian Institute of Information Technology. He received a PhD in Computing Science specialising in Human Computer Interaction (HCI) and Machine Learning (ML) from University of Glasgow, Scotland. Research Areas: Exploratory Data Analysis, Time Series and Spatial Analysis Human Computer Interaction (HCI), Machine Learning, Digital Signal Processing, Digital and Analogue Electronics, Embedded System, and Sensors.