Criteria Redesign for Student Loan Consideration Using Factor Analysis and Data Clustering Approach

Klangwaree Chaiwut[1], **Worasak Rueangsirarak**[2], and **Roungsan Chaisricharoen**[3]

ABSTRACT

This paper presents the importance of redesigning the student loan consideration criteria which had been revealed to have some fault in evaluating the candidates. The historical data of student loan candidates elicited from their application form in the 2016 academic year was collected and analyzed by using Factor Analysis. There are 507 samples with 17 information attributes. The factor analysis reduced the dimensions of the variance in the samples by identifying the discriminative factors for student loan consideration. The experimental result shows that only nine factors were identified as discriminative factors, which are 1) Part-time job taken by the student, 2) Other scholarships that the student had been receiving, 3) Father's salary, 4) Family ownership of the land, 5) House rental expense, 6) Number of siblings in the family, 7) Number of siblings currently studying, 8) Amount of money that the student get from other scholarships, and 9) Parental Marital Status. The clustering technique was used to measure the group of important factors reduced from the factor analysis. The clustering result showed that the clusters are obviously separated from each other. Therefore, these discriminative factors were elicited by using factor analysis which can be used to reconstruct the student loan consideration criteria and implement a decision support system.

Keywords: Student Loan Consideration, Factor Analysis, Principal Components Analysis, Data Clustering, Euclidean Distance, Manhattan Distance, Criteria Redesign

1. INTRODUCTION

Education is a process used to facilitate learning that can help people have knowledge and make better changes in their lives. Education is one of the basic human needs. It can change human life to a state of positive wellness. It also enhances the knowledge, skills, and intelligence of the people. At the higher of education level, knowledge is organized for discovering, creating, capturing, sharing, and motivating students for enhancing their personal development. Higher education can be interpreted as reaching a higher quality of life [1]-[3]. Thus, the student who had been educated at the higher education level not only receives a higher future of revenues but is also opening many doors for a professional opportunity in their future [4].

It is the belief that a fair and equal educational opportunity is the reason for making society more equitable. The student loan is an alternative way to bring the educational opportunity for many students. The main objective of this student loan fund (SLF) is to financially support the students to enrol in the educational system; these targeted students are from low-income families. It can reduce the existing gap between a highly wealthy person and a low income status individual in terms of education. This SLF will support the tuition fees and living expenses during the students undergoing their learning program [5]-[8].

Student loan holds the key to providing many equal educational opportunities. There are several countries that select student loans policies as an important issue for developing a higher education strategy. Governments in many countries provide the SLF to their students; e.g. New Zealand, Chile, South Africa, Ethiopia, Hungary, Australia, China, Hong Kong, India, Indonesia, Japan, Korea, Malaysia, Philippines, Singapore, and Thailand [9], [10]. However, the higher level of education is facing an insufficient budget which is affected by the global financial crisis. Therefore, it cannot be a "one size fits all" philosophy in the financial scheme because each funding system cannot be granted with the uniform standards [11]. Since the budget for a student loan fund is limited, the selected student must pass an interview process from the committee which is a consideration of the student loan for the sake of complying with the policy of the loan's purpose. However, there is no efficient criterion for considering a poor student in the interviewing process which lead to the situation that some students are not able to get a fair opportunity in finishing up their education [9], [12].

Consequently, the consideration of student's quali-

Manuscript received on August 28, 2018 ; revised on February 11, 2019.

Final manuscript received on February 24, 2019.

[1,2,3] The authors are with School of Information Technology, Mae Fah Luang University, Chiang Rai, Thailand., E-mail: 5771501003@lamduan.mfu.ac.th, worasak.rue@mfu.ac.th and roungsan.cha@mfu.ac.th

[1] The author is with Student Affair department, University of Phayao, Phayao, Thailand., E-mail: klangwaree.ch@up.ac.th

[2] Corresponding author.

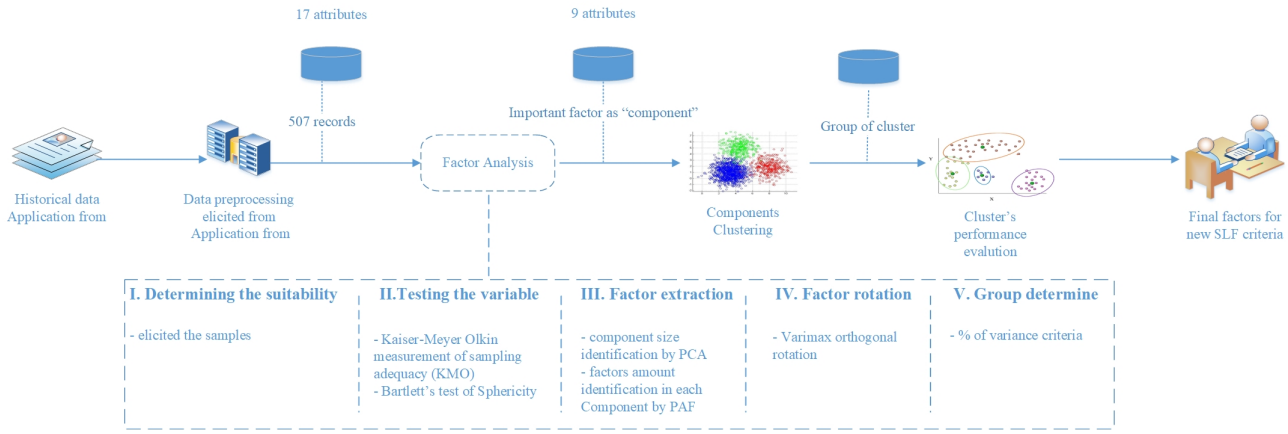

Fig.1: *Research Framework.*

fication is very important for making the student loan programs more effective. It could screen the applicants who really need financial support. In Thailand, the criteria in student loan consideration had been designed by the educational governor for more than ten years. However, the previous work by Klangwaree et al. used four data clustering techniques to identify the fault in the historical data of student loan, that have been considered as the result, with a different number of clusters scheme. The result shows a mix between accepted and rejected students as the members in each cluster. This means that there are some impoverished students who have lost the chance to receive a loan while some affluent students gain an easy access to the loan [12], [13]. This unfortunate situation is caused by some ineffective factors in the consideration criteria. Therefore, the student loan should be urgently revised and improved.

2. LITERATURE REVIEW

2.1 Student loan fund and its consideration system

In Thailand, the student loan fund was established in 1996 and supported by the government, which aims to financially support education for students coming from a poor income family. It will reduce the existing gap between rich and poor in terms of educational opportunity. The Thai student loan fund has provided loans for the student whose families' earn less than 20,000 Baht/year. The amount of grant includes tuition fees and living expenses during their study. The range of tuition fee is 15,000 - 60,000 Baht/year and the living expenses is about 2,400 Baht/month. The student repayments must be done after graduating for two years until 15 years. The interest rate is only one percent on the principal. The student loans budget and the quota of student allocated to the institution is limited. Thus, the selection method for a student to receive the student loan must meet the loan's purpose [10], [12], [15], [16].

In general, the student has to submit their request for a student loan fund via the E-Student loan System and send the document to the university through the faculty. After that, the committees will review the document and interview the student based on the structured criteria provided for committees. Normally, the committee members consist of a representative from many faculties within the university. Then, the candidates for a loan are selected based on the score evaluated from an interview order. However, there are many poor students that had not been selected to get the loan because the committee's opinion during the interviewing process is highly subjective. Therefore, the prejudices and personal preferences also might have distorted the consideration result [17]-[19].

2.2 Data analytic for screening applicants

Data analytic is a process of compiling and analyzing data with the goal of discovering useful information, informing conclusions, and supporting decision-making. The approach of data analytic depends largely on the type of data available for analyzing and the purpose of the analysis. Data analysis allows for the evaluation of data through analytical and logical reasoning to lead to some sort of outcome or conclusion in some context. It is a multi-faceted process which involves a number of steps, approaches, and diverse techniques [20]. Data analytics methodologies can be distinguished to Exploratory Data Analysis (EDA), which is the process on finding the patterns and relationships within the data, and Confirmatory Data Analysis (CDA), which is used together with statistical tools to identify the expected results from the data. In this study, we used the data analytic to explore how the information of the loan effect the applicants' data.

2.3 Factor analysis

Factor analysis is a method for reducing the information that have a similar pattern of response to get a small set of variable from a large dataset of information and summarizing the data. The purpose is so that the relationship and pattern can be easily interpreted. There are two types of factor analysis which are defined as Exploratory Factor Analysis (EFA), and Confirmatory Factor Analysis (CFA). EFA is used to determine the relationship between a variable and a factor, whereas CFA is used to measure the relationship among the factors. Factor analysis, like the clustering analysis, groups similar patterns variables into dimensions as reducing dimensionality. The factor analysis has been used in several fields such as behavioural and social sciences, medicine, economics, and geography [21]-[25]. In this study, we use Exploratory Factor Analysis (EFA) to analyse the data for finding the relation within it.

3. MATERIAL AND METHODS

In this research, the factor analysis and data clustering will be used to analyze the result of the traditional student loan consideration. The research framework is shown in Fig 1.

The algorithm of this research contains four stages: (1) data cleansing and preparation (2) factor analysis (3) data analysis using clustering techniques, and (4) evaluation of clustering results. Data cleansing is the process of detecting and removing errors from a data set. The factor analysis will group the attributes that have similar values to be a small set of parameters. In the third process of the research methodology, an evaluation of extracted factors by the factor analysis was tested using the clustering technique, K-mean, with the data set. The results of clustering will be evaluated to confirm the groups of clusters that is obviously separated from each other; these extracted factors can be used to distinguish the student loan fund candidates into two groups of the selected or non-selected students for loaning.

3.1 Data Pre-processing

There are two steps of preparation. The first stage is to investigate and impute the missing values. In the second stage, the missing values will be replaced with a zero value. In this study, the data padding will be used to process the data.

3.2 Factor Analysis

Factor analysis is a statistical approach that incorporate a similar pattern or dimension to uncover the latent dimensions of a variable. It reduces the multi-dimensional data into a low dimensional data [26]-[28]. Jianwei Nui et al. compared the application of data mining between the principal component analysis and factor analysis. The data is an anthropometric survey used by the Chinese National Institute of Standardization. Both of the algorithms were used to reduce the complexity of variances in the data set. The result shows that these techniques have both of the same and different variances. To select the right techniques, the domain expertise and the statistical pattern of the data are needed to be considered [28]. Wang Lijuan and Xu Ye used the factor analysis to identify the industry benefit. This concept can simplify the data structure and enhance industry competitiveness [29]. Gao Yarong and Guo Jianxiao used factor analysis to integrate an evaluation of key provincial disciplines. It grouped the reasonable indicators together. These results can be used to design the objective and provide a good method that is suitable for discipline estimation [27]. The process of factor analysis can be explained with the five steps of dimensional reduction [30] as follows:

Step 1: Determining the suitability of data for factor analysis
Step 2: Testing the variable using Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) and Bartlett's test of Sphericity.
Step 3: Factor extraction
Step 4: Factors rotation
Step 5: Group determine

The factor analysis model can be described as follow:

$$\begin{aligned} X_1 &= a_{11}F_1 + a_{12}F_2 + a_{13}F_3 + \ldots + a_{1m}F_m + a_1U_1 \\ X_2 &= a_{21}F_1 + a_{22}F_2 + a_{23}F_3 + \ldots + a_{2m}F_m + a_2U_2 \\ &\ldots \\ X_n &= a_{n1}F_1 + a_{n2}F_2 + a_{n3}F_3 + \ldots + a_{nm}F_m + a_nU_n \end{aligned} \tag{1}$$

where; X is an observed variable
i.e. $X = X_1, X_2, \ldots, X_n$
a_{ij} is a factor loadings
i.e. $a_{ij} = a(i = 1, 2, \ldots, n; j = 1, 2, \ldots, m)$
F is a common factor
i.e. $F = F_1, F_2, \ldots, F_m$
U is an unique factor
i.e. $U = U_1, U_2, \ldots, U_n$

Step 1: Determining the suitability of data for factor analysis

The historical data was elicited from the student loan candidates' application forms in the year 2016 that contained 507 samples. These factors are the discriminative information that can be represented as some critical issues that had an effect on the candidates' financial status. This is explained in table 1.

Step 2: Testing the variable using Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) and Bartlett's test of Sphericity

Kaiser-Meyer-Olkin Measurement of Sampling Adequacy (KMO) is a measurement on the appropriateness of the respondent data which will be used in the factor analysis. This test will be used to assess the adequacy of correlation matrices for the factor analysis [28], [30], [31].

$$KMO = \frac{\sum r_{i=j}^2}{\sum r_{ij}^2 + \sum u_{ij}} \tag{2}$$

where; $R = [r_{ij}]$ is the correlation matrix.
$U = [u_{ij}]$ is the partial covariance matrix.

KMO values between 0.8 and 1 indicate the sampling is adequate. KMO values less than 0.6 indicate the sampling is not adequate and that remedial action should be taken. KMO Values close to zero means that there are large partial correlations compared to the sum of correlations. However, some researches reduced the indicator to be 0.5 for identify inadequate of the sampling with limit of data [32], [33].

The Bartlett's Test of Sphericity is an assumption that the correlation matrix of different variables is not related which lead to an inappropriate implementation of the detection. It is an identity matrix that measures a relation edge of the indicator through other variables. After the measurement, it compares the observed correlation matrix to the identity matrix.

The Bartlett's Test of Sphericity value that has approximate Chi-square with a significant less than 0.05 will be indicated as the independent factors.

Step 3: Factor extraction

Factor extraction can be distinguished to be 2 subprocesses which are Principal Components Analysis and Principal Axis Factor analysis [34]. Principal components analysis (PCA) is to simply reduce the high dimension values into a smaller complexity dimension. The goal is to create a set of components while, Principal axis factor analysis (PAF) is recommended to use on the data that violate the assumption of multivariate normality. The goal of PAF is to create a list factor within each component [35]. If p is the variables $X_1, X_2, \ldots,$ then X_p is measured on a sample of i subject. Each factor will be estimated as the weighted sum of the p variables. For the i^{th} factor, the extraction can be explained as the following.

$$\begin{aligned} F_1 &= W_{(1)1}X_1 + W_{(1)2}X_2 + \ldots W_{(1)p}X_p \\ &\ldots \\ F_i &= W_{(i)1}X_1 + W_{(i)2}X_2 + \ldots W_{(i)p}X_p \end{aligned} \tag{3}$$

Step 4: A rotation processes

The purpose of rotation is to distinguish the discriminative factors from the data. It can identify a difference in the relevance of value within the factors based on its associations [36]. The rotation is used to simplify a structure and seeks to provide a more interpretable outcome. The rotation will maximize high item loadings and minimize low item loadings. Therefore, producing a more interpretable and simplified solution [31].

The orthogonal rotations are a rotation matrix that assumes the factors are not correlated or the factors are rotated 90° from each other. It is designed to produce a new set of approximating simple structures [35]. The components of the orthogonal are Varimax, Quartimax, and Equamax. The Varimax will simplify the values of factors and minimize the high loadings of them. Quartimax will identify the observed variables and reduce the factors that are used to explain each variable. Consequently, Equamax will combine the above methods to minimize highly loaded variables that influence the factors together. In this study, the orthogonal rotation based on Varimax was selected to rotate the factors [36]. The Varimax searches for a rotation of the original factors such that the variance of the loadings is maximized, which can calculate the amounts to be maximized (v) as follows;

$$v = \sum (q_{j,l}^2 - q_{j,l}^{-2})^2 \tag{4}$$

where; $q_{j,l}^2$ is the squared loading of the j^{th} variable on the l factor.
$q_{j,l}^{-2}$ being the mean of the squared loadings.

Step 5: Determining the group

In this process, the variables will be grouped based on the correlation. More than one variable can be grouped into the same component, in order to represent the high correlation between the variables in the component rather than the other components.

3.3 K-mean clustering

K-mean clustering is a popular and simple algorithm that has been proposed from 50 years ago. K-means can group each data point to a member of multiple clusters with a membership [37]. In this research, the K-mean will be used to categorize the factors reduced from the factor analysis. Therefore, the algorithm procedure can be described as the following;

Step 1: Design number of clusters, k, which are the points represented in the initial group centroids.

Step 2: Group the variables into the cluster with the nearest centroid, by using this equation:

$$Ex = \sum_{i=1}^{k} \sum_{p \in c_i}^{n} dist(p, c_i)^2 \tag{5}$$

where; E is the sum squared error for all objects in the data set.

n is a number of data.
k is a number of clusters.
p is the point in space representing a given object.
c is a centroids value

Step3: Compute the positions of the k centroids.
Step4: Repeat *Step 2* and *3* until the centroids do not change.

3.4 Evaluation of clustering results

In this study, five clustering evaluation techniques were used to identify the quality of the discriminative factors elicited by the factor analysis method.

3.4.1 Euclidean Distance

The Euclidean distance computes the length of the line segment between two points. It is regarded as the famous-used distance function [38].

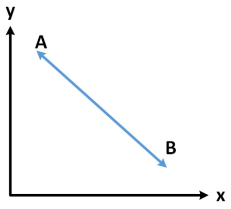

Fig.2: *Concept of Euclidean distance.*

From Fig.2, the distance between a pair of two points can be derived as the following equation:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{6}$$

where; n is the number of variables or points in the cluster.
x_i and y_i are the values of the i^{th} pairwise.

3.4.2 Manhattan Distance

The Manhattan Distance computes the sum of the absolute difference of two points' coordinates

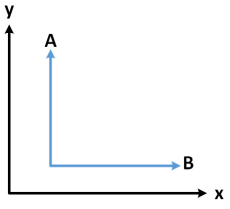

Fig.3: *Concept of Manhattan Distance.*

The calculation of the distance in *Fig.3* can be derived as the following equation:

$$d = \sum_{i=1}^{n} |x_i - y_i| \tag{7}$$

where; n is the number of variables or points in the cluster.
x_i and $y_i$are the values of the i^{th} pairwise.

3.4.3 Number of Iterations

The number of iterations is the method to measure the performance of the k-means algorithm in executing the data. The often iteration that must be stopped in this number for many times, even if the convergence criterion is not satisfied, will be selected. The parameters for evaluation are the number of iterations (NOI) which counts the number of iterations of K-Means to arrive at the convergence criteria: the sum of squares error [25], [26].

3.4.4 Sum of within-cluster distances

The sum of the within-cluster is the most common technique for partitioning a dataset of K-mean clustering using a non-hierarchical clustering procedure [39].

3.4.5 Within-cluster sum of squared

This method is applied to consider the summation of the squared differences between each observation and its group's average. It can be used as a measurement of variation within a cluster. It is the distances between a collection of points and the centroid associated with those points. This sum squared is equal to the sum of the pairwise squared distances between those points divided by the size (number of objects) of the collection [25], [27].

$$\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}(x_{ij} - \bar{x}_{kj})^2 \tag{8}$$

where; S_k is the set of observations in the k^{th} cluster
$\bar{x}_{kj}$ is the j^{th} average of the cluster center for the k^{th} cluster.
p is the number of clustering variables (dimensions)
X_{ij} is the value of variable p in cluster k

4. EXPERIMENTAL ANALYSIS

The proposed algorithm was applied to redesign the criteria for considering the student loan applicants. This study used the factor analysis to extract the discriminative information within the historical data of student loan candidates' application forms, which was partly performed with the SPSS Statistics 17.0. Then, an evaluation of extracted factors from the factor analysis was tested with the clustering technique, K-mean, using the Weka 3.6.13 which had been developed at the University of Waikato, New Zealand [42] with the number of clusters as 2, 4, 8 and 16.

Table 1: *Information attributes elicited from the application form.*

Attribute code	*Attribute description*	*Data Example*
X_1	Health status of the student	0 = healthy 1 = not healthy
X_2	A part-time job that is occupied by the student	0 = no job 1 = have a job
X_3	Other scholarships that the student had been receiving	0 = gets some other scholarships 1 = no scholarship
X_4	Amount of money that the student gets from other scholarships	2,500 Baht/month
X_5	Parental Marital Status	1 = married 2 = divorced 3 = father or mother died 4 = father and mother died
X_6	Father's salary	7,000 bath/month
X_7	Mother's salary	10,000 Baht/month
X_8	Parent's salary (third person)	7,000 Baht/month
X_9	Ownership of the land that the family is living on.	0 = their own land 1 = by rent or it belongs to others
X_{10}	House rental expense	5,000 Baht/month
X_{11}	Land size (Area in Rai unit)	3 Rai
X_{12}	Amount of debt within the family	5,000 Baht/month
X_{13}	Medical expense	9,000 Baht/month
X_{14}	Number of siblings in the family	3 children
X_{15}	Number of siblings who are studying	2 children
X_{16}	The highest education level of the sibling	0 = null 1 = Pre-elementary and less than, 2 = high school, 3 = diploma, and 4 = bachelor and higher
X_{17}	Monthly expense by student	3,000 Baht/month

Table 2: *KMO and Bartlett's Test.*

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.539
Bartlett's Test of Sphericity	Approx. Chi-Square	1599.103
	df	136.000
	Sig.	0.000*

*The significant < 0.0001

Table 2 shows the Kaiser-Meyer-Olkin (KMO) and Bartlett's Test of Sphericity. The KMO was used to assess the suitability of the respondent data for the factor analysis. In this experiment, the score of KMO is 0.539, which is less than 0.6. However, in Bartlett's Test of Sphericity, the hypothesis of the correlation matrix is used as an identity matrix, which would indicate that the variables are unrelated and, therefore, unsuitable for structuring a detection. The result from the SPSS shows that the Chi-Square is 1,599.103 (significance = ****), which is less than 0.05 of the significance level. It rejects the null hypothesis, thus indicating that 17 variables are related and suitable for the factor analysis. Therefore, in this study, we will use this limit data to analyze further steps, even the KMO is less than 0.6.

Table 3 shows the total variance that explains the model. The components are the same as the number of variables used in the factor analysis. In this research, there are 17 components which are equal to 17 variables.

The Initial Eigenvalues were calculated from the Principal Component Analysis in which the commonalities are one. Based on the result in table 3, the first seven components are meaningful as they have Eigenvalues greater than 1. Hence, these 7 components were selected to find the extraction sum for the further steps. It explains more variance than a single observed variable. The percent of total variance that had been accounted for each component can be described as the equation below.

$$\%\ of\ variance\ component_i = \frac{(Total_i \times number\ of\ variables)}{100} \quad (9)$$

where; $Total_i$ is the total variance for each i^{th} component.

The cumulative of the factor contains the cumulative percentage of variance accounted by the current and all preceding factors. The cumulative contribution rate of the seven components shows 63.38%, which is totally acceptable. The extraction sums of squared loadings were calculated from the extracted factors which is in the same way as the Initial Eigenvalues. After the rotation of the Varimax, the rotation sums of squared loadings will show the distribution of the variance. A high absolute value extracted from seventeen variables shows the influence of the factor to the loading variables. In table 3, the empty

Table 3: *Total Variance Explained.*

Components	*Initial Eigenvalues*			*Extraction Sums of Squared Loadings*			*Rotation Sums of Squared Loadings*		
	Total	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	2.017	11.867	11.867	2.017	11.867	11.867	1.877	11.042	11.042
2	1.906	11.215	23.082	1.906	11.215	23.082	1.835	10.794	21.837
3	1.785	10.498	33.580	1.785	10.498	33.580	1.802	10.602	32.439
4	1.718	10.106	43.686	1.718	10.106	43.686	1.684	9.906	42.345
5	1.190	6.999	50.685	1.190	6.999	50.685	1.337	7.862	50.207
6	1.115	6.560	57.245	1.115	6.560	57.245	1.156	6.801	57.008
7	1.044	6.139	63.384	1.044	6.139	63.384	1.084	6.375	63.384
8	0.978	5.751	69.135						
9	0.933	5.491	74.626						
10	0.894	5.257	79.883						
11	0.768	4.515	84.398						
12	0.723	4.253	88.650						
13	0.639	3.759	92.409						
14	0.472	2.779	95.189						
15	0.355	2.089	97.278						
16	0.278	1.636	98.913						
17	0.185	1.087	100.000						

Table 4: *Component Matrix.*

Factors	*Components*						
	1	2	3	4	5	6	7
X_6	0.608	-0.168	0.439	-0.216	0.097	0.040	-0.007
X_{14}	-0.600	0.122	0.595	-0.044	0.147	-0.204	-0.131
X_{10}	0.511	0.015	0.403	0.161	0.268	0.037	0.112
X_9	0.496	-0.104	0.422	0.033	0.259	-0.124	0.041
X_3	0.253	0.876	0.087	0.035	-0.169	0.038	-0.021
X_2	0.290	0.861	0.132	0.053	-0.159	0.010	0.018
X_7	-0.318	0.432	-0.388	0.169	0.250	-0.137	0.108
X_{15}	-0.557	0.130	0.656	-0.008	0.178	-0.200	-0.068
X_4	0.024	0.102	-0.122	-0.861	0.202	0.042	0.040
X_5	0.020	-0.146	0.127	0.848	-0.200	-0.034	-0.092
X_{12}	0.007	0.189	-0.304	0.198	0.508	-0.320	0.097
X_{16}	-0.142	0.019	0.294	-0.082	-0.500	0.213	0.354
X_1	0.057	0.003	0.032	0.261	0.470	0.415	0.045
X_8	-0.361	0.212	0.203	-0.048	0.149	0.522	-0.011
X_{11}	-0.254	-0.139	0.097	0.036	0.068	0.343	0.292
X_{17}	0.035	0.071	-0.133	0.067	0.103	0.485	-0.645
X_{13}	-0.023	0.049	-0.114	0.165	0.156	0.215	0.589

Table 5: *Rotated Component Matrix.*

Factors	*Components*						
	1	2	3	4	5	6	7
X_3	0.930	0.023	-0.005	-0.025	0.018	-0.055	0.047
X_2	0.928	0.083	0.002	-0.004	0.019	-0.047	0.001
X_6	-0.002	0.753	-0.083	-0.156	-0.201	-0.100	0.015
X_9	0.008	0.711	0.039	0.038	0.066	-0.052	-0.069
X_{10}	0.139	0.693	-0.012	0.124	0.070	0.122	-0.007
X_{15}	0.004	0.024	0.909	0.030	-0.025	0.038	-0.060
X_{14}	-0.019	-0.054	0.898	0.002	-0.030	-0.016	-0.022
X_4	-0.004	0.005	-0.021	-0.899	-0.005	-0.042	0.007
X_5	-0.030	0.035	-0.002	0.897	-0.009	0.001	0.034
X_{12}	0.047	-0.022	-0.026	0.017	0.719	0.048	-0.122
X_{16}	0.137	-0.091	0.089	0.037	-0.614	0.228	-0.265
X_7	0.251	-0.435	0.060	-0.001	0.525	0.157	-0.062
X_{13}	0.049	-0.014	-0.163	0.040	0.107	0.593	-0.259
X_{11}	-0.161	-0.070	0.100	0.002	-0.157	0.487	-0.005
X_1	-0.035	0.214	-0.016	0.106	0.243	0.485	0.339
X_8	0.127	-0.133	0.321	-0.117	-0.143	0.462	0.359
X_{17}	0.039	-0.067	-0.089	0.031	-0.002	-0.062	0.819

Table 6: *Components' members.*

Components	*List of members*	*Factor details*
Component #1	X_2	A part-time job that is occupied by the student
	X_3	Other scholarships that the student had been receiving
Component #2	X_6	Father's salary
	X_9	Ownership of the land that the families are living on.
	X_{10}	House rental expense
Component #3	X_{14}	Number of siblings in the family
	X_{15}	Number of siblings who are studying
Component #4	X_4	Amount of money that the student get from other scholarships
	X_5	Parental Marital Status
Component #5	X_{12}	Amount of debt within the family
	X_{16}	The highest education level of the sibling
	X_7	Mother's salary
Component #6	X_{13}	Medical expense
	X_{11}	Land size (Area in Rai unit)
	X_1	Health status of the student
	X_8	Parent's salary (third person)
Component #7	X_{17}	Monthly expense by student

cells represent the loadings which are less than 0.5. Consequently, the calculation of the rotated component matrix will be conducted in order to identify the important variables which will be selected to be a member of the component. This is shown in table 4.

In table 5, the rotated factor loadings values are calculated by reducing the number of factors who have a high loading of variables. The proper coefficient value to identify loading in the rotated factor are the values between -1 and 1. We find that seven components can be extracted as shown in table 6.

In table 6, the first common factor (Component #1) has more loading on X_2-*other scholarships that the student had been receiving* and X_3-*a part-time job that is occupied by the student*, with the Eigenvalue of 2.017. The second common factor (Component #2) has more loading on X_6-*father's salary*, X_9-*ownership of the land that the family is living on* and X_{10}-*house rental expense*, with the eigenvalue of 1.906. The third common factor (Component #3) has more loading on X_{14}-*the number of siblings in the family* and X_{15}-*the number of siblings who are studying*, with the Eigenvalue of 1.785. The fourth common factor (Component #4) has more loading on X_4-*Amount of money that the student gets from other scholarships* and X_5-*Parental Marital Status*, with the eigenvalue of 1.718. The fifth common factor (Component #5) has more loading on X_{12}-*the amount of debt within the family*, X_{16}-*highest education level of the sibling* and X_7-*mother's salary*, with the Eigenvalue of 1.190. The sixth common factor (Component #6) has more loading on X_{13}-*the medical expense*, X_{11}-*land size*, X_1-*the health status of the student* and X_8-*parent's salary*, with the Eigenvalue of 1.115. And the seventh common factor (Component #7) has more loading on $X_{17}-$*monthly* expense by the student, with the Eigenvalue of 1.044.

In the final process of this research methodology, an evaluation of extracted factors by factor analysis was tested using the clustering technique, K-mean [14] with two different distance functions (Euclidean and Manhattan) on $k = 2, 4, 8$ and 16. The extracted factors consisted of two datasets from the factor analysis result. We distinguished the selected components to be two sets of components which are the components that have a cumulative % of variance $\leq$50 % and the components that have a cumulative % of variance between 11.87- 63.38%. As a result, there are four components that were selected to the 1st set of data and seven components were selected for the 2nd set of data which are nine factors from Component #1 to Component #4 and seventeen factors from Component #1 to Component #7. The result can be shown in table 7, 8, and 9. Therefore, these two groups of factors (from difference amount of component) were selected based on the value of % of Variance in table 3; which are % *of Variance* $>$ *10* for "4 components", and % *of Variance* $>$ *6* for "7 components".

Table 7: *Number of iterations.*

Number of clusters	*Euclidean distance*		*Manhattan distance*	
	4 components	7 components	4 components	7 components
2	2	3	2	3
4	5	11	3	5
8	6	15	5	8
16	9	9	5	9

Table 7 shows the number of iterations. In the different number of clusters with both distance measurements, the number of iterations of four components is less than seven components. It shows that four components perform better than seven components because the computational time complexity is

proportional to the number of iterations.

Table 8: *Sum of within-cluster distances evaluation.*

Number of clusters	*Manhattan distance*	
	4 components	*7 components*
2	581.36	785.24
4	457.85	594.33
8	338.31	527.04
16	219.27	436.45

Table 8 shows the sum of within-cluster distance value which is evaluated by using only the Manhattan distance. This is the distance between each member and the centroid. The distance of four components is shorter than seven components. The shortest distance, 219.27 on four components, represents the highest density and best performance to analyze the student loan criteria by using K-mean [38], [43].

Table 9 shows the within-cluster sum of squared errors which is evaluated by using the Euclidean distance. The result shows that the sum of squared errors of the four components is less than seven components. The sum of squared errors was computed by each instance in the cluster by summing the squared differences between each attribute value and the corresponding one in the cluster centroid. It can be used as a measurement of variation within a cluster [38], [43].

Table 9: *Within cluster sum of squared errors Evaluation.*

Number of clusters	*Euclidean distance*	
	4 components	*7 components*
2	293.13	371.79
4	151.32	256.48
8	114.03	166.67
16	61.24	120.82

Furthermore, this study has asked the student loan fund experts to evaluate the new set of factors to validate the applicant of this loan by completing a structured questionnaire adapted from [44]. The experts have to express their satisfaction based on the new criteria design for measuring new student loan applicants. There are five levels of ratings score as the standard of Linkert scale. The satisfaction levels can be distinguished to be "Strongly Disagree" for the score between 0.01 - 1.00, "Disagree" for the score between 1.01 - 2.00, "Undecided" for the score between 2.01-3.00, "Agree" for the score between 3.01-4.00, and "Strongly Agree" for the score between 4.01-5.00, respectively. The results from five invited experts can be shown as table 10.

In table 10, the experts agree on most of the questions with the average scores at 4.21 out of 5. This means that the new criteria are proper to evaluate the students who want to get the scholarship. However, in some items e.g. "how to assess applicant in an interview", and "accessing of the applicant background" are rated as fair level (3.01-4.00). This is because the new criteria have never been used to be evaluated on the large amount of applicants by them. Therefore, the experts are not sure about the performance of the criteria on this perspective. Hence, the validation of the new criteria with the real student loan consideration should be conducted for future works.

Table 10: *Item Recommended for Users by TD.*

Item for measuring new student loan fund criteria	*Average*	*S.D.*
1. I understand the criteria easily.	4.75	0.46
2. I can use the new criteria for my decision making.	4.12	0.64
3. I have the ability to design whether the student loan is in financial problem, or not.	4.00	0.75
4. I know how to assess the applicant in an interview.	3.87	0.64
5. I can access the student's . background in multiple dimensions	3.75	0.71
6. I can use the new criteria as the way for selecting.	4.25	0.87
7. I believe in the validity and reliability of the new criteria.	4.25	0.46
8. I know the purpose and objective for the new criteria.	4.62	0.52
9. I know how and what to assess on the candidate.	4.37	0.52
10. I can evaluate and select the candidate based on their appropriateness to get a scholarship in a new criteria.	4.12	0.64
Total	4.21	0.14

Table 11: *Statistics of salary data from 507 samples.*

Type of salary	Number of available data	Number of unavailable data	Averaged value (THB)	S.D.
Father's salary	427	80	5845.52	5485.44
Mother's salary	167	340	4754.79	9155.42
Parent's salary	144	363	8048.13	18227.08

5. CONCLUSION

The factor analysis reveals the discriminative variables which are summarized to be only nine factors from Component #1 to Component #4. These discriminative factors can influence the student loan consideration within the clusters of the candidates' data using K-mean to be categorized with two different distance functions (Euclidean and Manhattan). Therefore, using the number of cluster $k = 16$ indicates that the shortest of the summation of within-cluster distance and the within-cluster sum of squared errors are in Component #1 to Component #4. These shortest distance and shortest summation of squared

errors represent the highest density, and confirm its best performance to analyze the student loan by using K-mean.

The factors in Component #1 to Component #4 consists of the *part-time job that is occupied by the student, other scholarships that the student had been receiving, father's salary, ownership of the land that the families are living on, house rental expense, number of siblings in the family, number of siblings who are studying, amount of money that the student get from other scholarships*, and *parental marital status*. On the other hand, the factors in Component #5 to Component #7 only have a little influence on the student loan consideration based on their Eigenvalues. For example, component #5 and component #6 contain the mother's salary and parents' salary which were not selected after the evaluation. This is because in our 507 samples, there are only 167 samples that have the mother's salary when their father had left the family. Also, there are only 144 samples that have the parents' salary when they had been left by their father or mother. With a high value of S.D. shown in table 11, it leads to not being influenced in our experiment.

This study also shows that the relevance of result only with factors in Component #1 to Component #4 are the most dominant factors to consider in providing the student loan fund based on the results of Eigenvalues which is greater than other components (#5 to #7). Hence, this experimental result can confirm the discriminate extracted factors through the K-mean clusters which lead to an accurate opportunity for the indigent students. In other words, the current criteria has only 9 discriminative factors out of 17 factors. Also, these 9 factors have been confirmed by the student loan fund experts that should have more efficiency than the current criteria (17 factors).

This work proclaims that there is an urgent need to review the student loan consideration criteria and reconstruct the criteria based on the most discriminative information from the raw data elicited from the application forms for candidate evaluation. It is also recommended that more data needs to be collected for the purpose of using the information as inquiries during the interview process by the committee. Altogether, these discriminative factors could help the committee to efficiently select the right student by applying a developed decision support system with a novel machine learning approach. Besides that, this system could be used for granting other scholarships to students as well. Also, these 9 factors have been confirmed by the student loan fund experts that should have more efficiency than the current criteria (17 factors).

ACKNOWLEDGMENT

This research was sponsored by Mae Fah Luang University. The authors of this paper are highly grateful to the University of Phayao staffs in providing the technical assistance.

References

[1] F. Simon, K. Malgorzata, and P. Beatriz, *Education and Training Policy No More Failures Ten Steps to Equity in Education: Ten Steps to Equity in Education*. OECD Publishing, 2007.

[2] H. Deveci, "Value Education through Distance Learning: Opinions of Students Who Already Completed Value Education," *Turk. Online J.Distance Educ.*, vol. 16, no. 1, pp. 112-126, Jan. 2015.

[3] S. Natex and D. Lesjak, "Integrated higher education information systems-professors' knowledge management tool," *Issues Inf. Syst.*, pp. 80-86, 2011.

[4] D. E. Bloom, M. Hartley, and H. Rosovsky, "Beyond private gain: The public benefits of higher education," in *International handbook of higher education*, Springer, 2007, pp. 293-308.

[5] W.-R. Shi, Y.-H. Lin, Y. Li, T.-N. Yang, and L. Zhang, "A matter-element evaluation model for the individual credit of university students loan," in *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, vol. 5, pp. 2853-2858, 2004.

[6] H. Wei, "An International Estimation on Multiattribute of Student Loan Scheme Based on Entropy Theory," in *Information Management, Innovation Management and Industrial Engineering, 2008. ICIII'08. International Conference on*, vol. 3, pp. 8-12, 2008.

[7] D. Zhao, "Integrating RFM model and cluster for students loan subsidy valuation," in *Business and Information Management, 2008. ISBIM'08. International Seminar on*, vol. 2, pp. 461-464, 2008.

[8] B. W. Ambrose, L. Cordell, and S. Ma, "The Impact of Student Loan Debt on Small Business Formation," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2633951, Jul. 2015.

[9] M. Woodhall, *Student Loans in Higher Education: 2. Asia: Report of an IIEP Educational Forum*. International Institute for Educational Planning, 1991.

[10] B. Chapman and K. Lounkaew, *11. Income-Contingent Student Loans for Thailand: Alternatives compared*. ANU E Press, 2011.

[11] V. Skrbinjek, D. Lesjak, and others, "Changes in higher education public funding during economic and financial crisis," in *V: Management, Knowledge and Learning International Conference*, pp. 25-27, 2014.

[12] Z. Zhao, W. Zhang, and Y. Zhou, "National student loans credit risk assessment based on GABP algorithm of neural network," in *Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on*, pp. 2196-2199, 2011.

[13] K. Chaiwut, W. Rueangsirarak, and R. Chaisricharoen, "Factor analysis on student loan consideration in higher education level," in *2017 International Conference on Digital Arts, Media and Technology (ICDAMT)*, pp. 296-301, 2017.

[14] K. Chaiwut, W. Rueangsirarak, and R. Chaisricharoen, "Fault Detection in Student Loan Analysis using Clustering Techniques.," *Int. Conf. Digit. Arts Media Technol.*, pp. 200-203, Mar. 2016.

[15] M. A. Salam, "THAI STUDENT LOAN FUND AND ITS CURRENT STATUS.," *J. Asia Pac. Stud.*, vol. 5, no. 1, 2018.

[16] A. Ziderman, "The student loan schemes in Thailand: A review and recommendations for efficient and equitable functioning of the scheme," *Bangk. U. N. Educ. Sci. Cult. Organ.*, 1999.

[17] I. Irvanizam, "Application of the fuzzy topsis multi-attribute decision making method to determine scholarship recipients," *J. Phys. Conf. Ser.*, vol. 978, no. 1, p. 012056, 2018.

[18] I. Irvanizam, "Multiple attribute decision making with simple additive weighting approach for selecting the scholarship recipients at Syiah Kuala university," in *Electrical Engineering and Informatics (ICELTICs), 2017 International Conference on*, pp. 245-250, 2017.

[19] G. A. M. S. Wimatsari, I. K. G. D. Putra, and P. W. Buana, "Multi-attribute decision making scholarship selection using a modified fuzzy TOPSIS," *Int. J. Comput. Sci. Issues IJCSI*, vol. 10, no. 1, p. 309, 2013.

[20] B. Selene Xia and P. Gong, "Review of business intelligence through data analysis," *Benchmarking Int. J.*, vol. 21, no. 2, pp. 300-311, 2014.

[21] A. G. Yong and Sean Pearce, "A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis," *Tutor. Quant. Methods Psychol.*, vol. 2013, no. 9(2), pp. 79-94.

[22] N. Wu and J. Zhang, "Factor analysis based anomaly detection," in *Information Assurance Workshop, 2003. IEEE Systems, Man and Cybernetics Society*, pp. 108-115, 2003.

[23] P. K. Sari, N. Nurshabrina, and Candiwan, "Factor analysis on information security management in higher education institutions," in *2016 4th International Conference on Cyber and IT Service Management*, pp. 1-5, 2016.

[24] J. Niu et al., "A comparative study on application of data mining technique in human shape clustering: Principal component analysis vs. Factor analysis," in *2010 5th IEEE Conference on Industrial Electronics and Applications*, pp. 2014-2018, 2010.

[25] B. Williams, T. Brown, and A. Onsman, "Exploratory factor analysis: A five-step guide for novices," *ResearchGate*, vol. 8, no. 3, pp. 1-13, Jan. 2010.

[26] N. Wu and J. Zhang, "Factor analysis based anomaly detection," in *Information Assurance Workshop, 2003. IEEE Systems, Man and Cybernetics Society*, pp. 108-115, 2003.

[27] G. Yarong, G. Jianxiao, P. Shanna, and S. Lei, "Integrated evaluation of key provincial disciplines in colleges and universities based on factor analysis method," in *Future Information Technology and Management Engineering, 2009. FITME'09. Second International Conference on*, pp. 278-281, 2009.

[28] J. Niu et al., "A comparative study on application of data mining technique in human shape clustering: Principal component analysis vs. Factor analysis," in *2010 5th IEEE Conference on Industrial Electronics and Applications*, pp. 2014-2018, 2010.

[29] W. Lijuan and X. Ye, "Factor analysis on industrial efficiency of fiscal supporting SMES technology innovation," in *2011 International Conference on Business Management and Electronic Information*, vol. 1, pp. 9-11, 2011.

[30] P. K. Sari and N. Nurshabrina, "Factor analysis on information security management in higher education institutions," in *Cyber and IT Service Management, International Conference on*, pp. 1-5, 2016.

[31] B. Williams, A. Onsman, and T. Brown, "Exploratory factor analysis: A five-step guide for novices," *Australas. J. Paramed.*, vol. 8, no. 3, pp. 1-131, 2010.

[32] H. F. Kaiser, "An index of factorial simplicity," *Psychometrika*, vol. 39, no. 1, pp. 31-36, 1974.

[33] B. A. Cerny and H. F. Kaiser, "A study of a measure of sampling adequacy for factor-analytic correlation matrices," *Multivar. Behav. Res.*, vol. 12, no. 1, pp. 43-47, 1977.

[34] K. B. Coughlin, "An analysis of factor extraction strategies: A comparison of the relative strengths of principal axis, ordinary least squares, and maximum likelihood in research contexts that include both categorical and continuous variables," 2013.

[35] A. G. Yong and S. Pearce, "A beginner's guide to factor analysis: Focusing on exploratory factor analysis," *Tutor. Quant. Methods Psychol.*, vol. 9, no. 2, pp. 79-94, 2013.

[36] H. Abdi, "Factor rotations in factor analyses," *Encycl. Res. Methods Soc. Sci. Sage Thousand Oaks CA*, pp. 792-795, 2003.

[37] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651-666, 2010.

[38] R. Loohach and K. Garg, "Effect of distance functions on simple k-means clustering algorithm," *Int. J. Comput. Appl.*, vol. 49, no. 6, 2012.

[39] D. Steinley and M. J. Brusco, "Initializing k-means batch clustering: A critical evaluation of several techniques," *J. Classif.*, vol. 24, no. 1, pp. 99-121, 2007.

[40] P. S. Bishnu and V. Bhattacherjee, "Software fault prediction using quad tree-based k-means clustering algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1146-1150, 2012.

[41] M. J. Brusco, "A repetitive branch-and-bound procedure for minimum within-cluster sums of squares partitioning," *Psychometrika*, vol. 71, no. 2, pp. 347-363, 2006.

[42] N. Sharma, A. Bajpai, and M. R. Litoriya, "Comparison the various clustering algorithms of weka tools," *facilities*, vol. 4, no. 7, 2012.

[43] Y. S. Thakare and S. B. Bagal, "Performance evaluation of K-means clustering algorithm with various distance metrics," *Int. J. Comput. Appl.*, vol. 110, no. 11, 2015.

[44] Z. Hosseini and A. Kamal, "Developing an instrument to measure perceived technology integration knowledge of teachers," in *Proceedings of International Conference of Advanced Information System, E-Education & Development*, pp. 7-8, 2012.

Klangwaree Chaiwut is currently a PhD candidate at School of Information Technology, Mae Fah Luang University, Thailand. She is a staff at the Student Affair department, University of Phayao, Thailand. She received B.Sc. degree in Computer Science in 2006, and M.Sc. degree in Internet and Information Technology in 2011 from Naresuan University, Thailand. Her research interests are Clustering Techniques, factor analysis, and machine learning. She was also a scholarship researcher of Institute of Information Technology and Doctoral School of Business Information at the Corvinus University of Budapest, Hungary.

Worasak Rueangsirarak is a lecturer at School of Information Technology, Mae Fah Luang University, Thailand. He has an MSc in Computer Science and a PhD in Knowledge Management. His background is computer science with special interests in Healthcare Technology, Applied Computing and AI, Human Motion Analysis, Knowledge Management, Knowledge Engineering, Expert System, Decision Support System, Computer Graphic and Animation, Computer Simulation and Visualization, Geographic Information System. He received the grants for working as a postdoctoral research fellow at Visual Computing research lab, Northumbria University, UK, and a visiting researcher based at Faculty of Computing, Science and Technology, Staffordshire University, UK.

Roungsan Chaisricharoen received the Ph.D. degree in 2009 from the Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand. He is an Assistant Professor of Computer Engineering at the School of Information Technology, Mae Fah Luang University, Thailand. He is now the chairperson of both the Master and Ph.D. programs in Computer Engineering. His research interests are computational intelligence, data communication, optimization, application of ICT in agriculture, embedded system, and analogue integrated circuit.