Indonesian Journal of Statistics and Its Applications
Available at https://ijsa.stats.id
e-ISSN: 2599-0802

K-prototypes Algorithm for Clustering Schools Based on The Student Admission Data in IPB University*

Sri Sulastri[1], Lismayani Usman[2], and Utami Dyah Syafitri[3‡]

[1,2,3]Department of Statistics, IPB University, Indonesia
[1,2]BPS-Statistics Indonesia, Indonesia
[‡]corresponding author: utamids@apps.ipb.ac.id

Abstract

The new student admissions were regularly held every year by all grades of education, including in IPB University. Since 2013, IPB University has a track record of every school that has succeeded in sending their graduates, even until they successfully completed their education at IPB University. It was recorded that there were 5,345 schools that included in the data. It was necessary to making every school in the data into the clusters, so IPB could see which schools were classified as good or not good in terms of sending their graduates to continue their education at IPB based on the characteristics of the clusters. This study using the k-prototypes algorithm because it can be used on the data that consisting of categorical and numerical data (mixed type data). The k-prototypes algorithm could maintain the efficiency of the k-means algorithm in handling large data sizes, but eliminated the limitations of k-means. The results showed that the optimal number of clusters in this study were four clusters. The fourth cluster (421 school members) was the best cluster related to the student admission at IPB University. On the other hand, the third cluster (391 school members) was the worst cluster in this study.

Keywords: clustering, k-prototypes, student admission.

* Received: Jan 2021; Reviewed: Jan 2021; Published: Jun 2021

1. Introduction

In every year, all grades of education routinely hold new student admissions, including state and private universities. This momentum is used as a starting point to finding the excellent student who truly deserve to be the part of the almamater. IPB University is the best tertiary education in Indonesia according to The Ministry of Education and Culture of Indonesia in 2020. Every year, IPB University has a track record of every school that has succeeded in sending their graduates to continue their education at IPB University. The data was recorded at the Directorate of Education, Administration and New Student Admissions of IPB. It was necessary to making every school in the data into the clusters, so IPB could see which schools were classified as good or not good in terms of sending their graduates to continue their education at IPB based on the characteristics of the clusters.

One of the popular and efficient algorithms that used in clustering large data sets is the k-means clustering algorithm. This algorithm partitioned data sets into k clusters that have been determined and calculated the center point of the clusters, so that a criterion function can be achieved (Anderberg 1973). However, this algorithm has limitations on numerical data types only, while the track record data that will be used in this study includes numerical and categorical attributes. Therefore, this research used the development of the k-means algorithm which can be used for clustering mixed data, which called k-prototypes algorithm.

K-prototypes algorithm can be used on data that consisting of categorical and numerical data. In general, the k-prototypes algorithm can maintain the efficiency of the k-means algorithm in handling large data, but eliminates the limitations of implementing of k-means. The limitations that referred before was that the k-prototypes algorithm can increase the homogeneity of data within the cluster and maximize heterogeneity between the clusters (Huang 1998).

Therefore, this study used the k-prototypes algorithm for clustering every school that have sent their graduates to IPB from 2013 to 2019. This study has two final purpose, the first was to obtain the optimal number of cluster that can describe the condition of schools based on their track record. Then the second purpose was to find out the characteristics of the cluster, both based on general conditions and based on its categorical and numerical variables.

2. Literature review

2.1 Cluster analysis

Cluster analysis is a method of clustering various data in the data set to become several clusters based on their respective characteristics, where the clusters that created will be homogeneous within the cluster and heterogeneous between the clusters (Liao 2005). This analysis belongs to the category of unsupervised learning that does not require an initial reference to obtain characteristics in making a cluster, where the results can be presented in description and visualization.

One of the benefits of this analysis is that it does not require the assumption of data distribution in its processing (Li et al. 2008). Cluster analysis is divided into two, that called hierarchical cluster analysis and nonhierarchical cluster analysis. The fundamental difference between both of them is about the formation of the number of clusters. The hierarchy method is through a process of agglomerative and devisive. The agglomerative processing of each object is considered as a separate cluster, then two clusters that have similarities are combined into a new cluster, and so the next step. Otherwise, the process of devisive starts from a large cluster consisting of all objects, then the highest dissimilar object is separated from the large cluster, and so on. As for the nonhierarchical method, it starts by determining the number of clusters that desired, then the clustering process can be runned without following the process that occurs in the hierarchical method.

2.2 K-prototypes algorithm

This algorithm is included in the nonhierarchical clustering method and was the first proposed by Huang (1998). The purpose of this algorithm is to making X data sets into k clusters by minimizing the cost function. This algorithm built on three processes, initial prototype selection, initial allocation, and reallocation. There are four steps of the k-prototypes algorithm:

Step 1: Determine the centroid of the cluster as many as the k clusters as the starting point C1, C2, ..., Ck on every variables {X1, X2, …, Xp};

Step 2: Calculate the distance of data points on the data set against the centroid of the cluster, then allocate the data points into the cluster that has the closest prototype distance with centroid;

Step 3: Calculate the new centroid of the cluster after all objects have been allocated into clusters, and then reallocate all objects on the new prototype;

Step 4: If the centroid of the cluster does not change or has been convergent, the algorithm would stop. However, if the centroid is still changing significantly, the process should return to step 2 and 3 until the maximum iteration is reached or there is no movement of the object.

2.3 Similarity measurement

Similarity measure is a measure of the resemblance of an object to a reference object. One of the commonly measure is using the distance based on similarity measure, where the similarity measure is calculated based on the distance measure. The greater distance between two objects, so become more different that two objects, and vice versa the smaller distance between the two objects, become more similar that object (Rencher 2007). Some types of distance measurements are Euclidean distance, Manhattan distance, Mahalanobis distance, categorical data type distance (simple matching) and mixed data type distance. In this study, we used the distance of mixed data types because the data on this study has two types data, such as categorical and numerical data.

Similarity measures of mixed data types proposed by Huang (1998) can be written in the following formula:

$$d(i,j)=\sum_{r=1}^{p}\left(x_{ir}-x_{jr}\right)^2+\gamma\sum_{s=p+1}^{m}\delta\left(x_{is},x_{js}\right) \tag{1}$$

with

$d(i,j)$: the distance of the i-th object with the j-th object (mixed variable)

$\sum_{r=1}^{p}\left(x_{ir}-x_{jr}\right)^2$: a measure of distance for numerical type data variables

$\gamma\sum_{s=p+1}^{m}\delta\left(x_{is},x_{js}\right)$: a measure of distance for categorical type data variables

x_{ir} : the value of the i-th object on the r variable

x_{jr} : the value of the j-th object on the r variable

p : the number of variables with numerical data types

$\delta(x_{is},x_{js})$: $\begin{cases} 0 \text{ when } x_{is} = x_{js} \\ 1 \text{ when } x_{is} \neq x_{js} \end{cases}$

m : the number of variables with categorical data types

According to Huang (1998), the value of gamma coefficient (γ) is obtained from the average standard deviation (σ) of all numerical variables. Based on the simulation conducted by Huang (1998), the value of gamma coefficient (γ) is obtained from 1/3 σ to p / 3 σ and the optimal gamma coefficient (γ) is commonly used by (p / 2) / 3 σ, with p is the sum of all variables. Therefore, the value of gamma coefficient (γ) is influenced by the number of objects (n), the number of numerical variables, and the number of categorical variables.

2.4 Evaluated the result of the clustering

The results of the clustering were evaluated using diversity values. If within the cluster is more homogeneous and between clusters is more heterogeneous, then the clustering shows the optimal results. It means that the diversity within the cluster is smaller and the diversity between the clusters is greater. The results of clustering can be measured using the ratio between the standard deviation between clusters (S_B) and the standard deviation within clusters (S_W). The smaller value ratio of S_W and S_B means that the clustering can be said to be optimal. The calculation for numerical type data can be written in the following formula (Bunkers et al 1996):

$$S_{wn}=\frac{1}{C}\sum_{C=1}^{C} S_C \tag{2}$$

with

S_{wn} = the standard deviation within cluster

S_C = the standard deviation of the c-th cluster

C = the number of clusters

The diversity between clusters can be calculated by the formula:

$$S_{Bn}=\left[\frac{1}{C-1}\sum_{C=1}^{C}(\bar{x}_c-\bar{x})^2\right]^{1/2} \tag{3}$$

with

S_{Bn} = the standard deviation between cluster of the numerical variables

$\bar{x}_c$ = c-cluster average value

$\bar{x}$ = the overall average value of the cluster

However, the diversity between clusters and within clusters for categorical data is calculated using a different formula. The following calculation formula according to Okada (1999) and Kader and Perry (2007):

$$S_{wk}=[MSW]^{1/2} \tag{4}$$

with

S_{wk} = the standard deviation within cluster of the categorical variables

MSW = mean of square within cluster

MSW can be obtained in the following formula:

$$MSW=\frac{SSW}{n-C}=\frac{1}{n-C}\left[\frac{n}{2}-\frac{1}{2}\sum_{C=1}^{C}\frac{1}{n_C}\sum_{k=1}^{K}n_{kc}^2\right] \tag{5}$$

with

SSW = sum of square within cluster
n = the number of observations
n_k = the number of k-th category observations (k = 1, 2, 3, …, K).
n_{kc} = the number of observations with the k-th category in the c cluster (c = 1 ,2, ..., C)
The total number of observations can be calculated with:

$$n=\sum_{C=1}^{C} n_c = \sum_{k=1}^{K} n_k = \sum_{k=1}^{K}\sum_{c=1}^{C} n_{kc} \tag{6}$$

The diversity values among clusters in the categorical data (S_{Bk}) can be calculated using the following formula:

$$S_{Bk}=[MSB]^{1/2} \tag{7}$$

with MSB is the mean of square between cluster that can be obtained by using the following formula:

$$MSB=\frac{SSB}{(C-1)}=\frac{1}{C-1}\left[\frac{1}{2}\left(\sum_{C=1}^{C}\frac{1}{n_C}\sum_{k=1}^{K}n_{kc}^2\right)-\frac{1}{2n}\sum_{k=1}^{K}n_k^2\right] \tag{8}$$

with SSB is the sum of square between cluster.

3. Methodology

3.1 Source of data

The data that used in this study is the trade record from 5,345 schools that their graduates accepted in IPB University from 2013 until 2019. This data is sourced from the database of the Directorate of Education Administration and New Student Admission, IPB University. There are 16 variables used in this study, that 12 variables are the categorical variables and 4 variables are the numerical variables (Table 1).

3.2 Method of analysis

There are four step of analysis carried out in this study. First step is the data preparation, then the clustering step, the cluster evaluation step and the results visualization step for the last step. The analysis and processing data is using R software.

1. The data preparation
   Checking the missing data and explore the data to get a descriptive statistics.
2. The clustering
   The clustering process in this study is using the k-prototypes algorithm by first determining the number of clusters (k) to be formed. The k limit according to Lin et al (2005) is minimum of 2 clusters and maximum of √n or n/2 clusters with n is the number of observations.
3. The cluster evaluation such as:
   a. calculate Swn in every numerical variable;
   b. calculate Swk in every categorical variable;
   c. calculate the total of Sw that was the average of Sw from every variable;
   d. calculate Sbn in every numerical variable;
   e. calculate Sbk Swk in every categorical variable;
   f. calculate the total of Sb that was the average of Sb from every variable;
   g. calculate the ratio of the total of Sw and the total of Sb.

4. Vizualization
   This step is to visualizing and interpreting the optimal clustering results in the form of graphs and tables, and to find out the characteristics of each cluster.

Table 1: List of variables

Variables	Type	Explanation
X1_2nd semester of 2013	Categorical	School’s predicate in 2nd semester of 2013
X2_1st semester of 2014	Categorical	School’s predicate in 1st semester of 2014
X3_2nd semester of 2014	Categorical	School’s predicate in 2nd semester of 2014
X4_1st semester of 2015	Categorical	School’s predicate in 1st semester of 2015
X5_2nd semester of 2015	Categorical	School’s predicate in 2nd semester of 2015
X6_1st semester of 2016	Categorical	School’s predicate in 1st semester of 2016
X7_2nd semester of 2016	Categorical	School’s predicate in 2nd semester of 2016
X8_1st semester of 2017	Categorical	School’s predicate in 1st semester of 2017
X9_2nd semester of 2017	Categorical	School’s predicate in 2nd semester of 2017
X10_1st semester of 2018	Categorical	School’s predicate in 1st semester of 2018
X11_2nd semester of 2018	Categorical	School’s predicate in 2nd semester of 2018
X12_1st semester of 2019	Categorical	School’s predicate in 1st semester of 2019
X13_The number of undergraduate student	Numerical	The number of school’s graduates who accepted and graduated from IPB University from 2nd semester of 2013 until 1st semester of 2019
X14_GPA ≥ 3.50	Numerical	The number of school’s graduates who graduated from IPB University with GPA ≥ 3.50 from 2nd semester of 2013 until 1st semester of 2019
X15_GPA ≥ 2.75 and < 3.50	Numerical	The number of school’s graduates who graduated from IPB University with GPA ≥ 2.75 and < 3.50 from 2nd semester of 2013 until 1st semester of 2019
X16_GPA < 2.75	Numerical	The number of school’s graduates who graduated from IPB University with GPA < 2.75 from 2nd semester of 2013 until 1st semester of 2019

4. Results and Discussion

4.1 Variables Description

Data from the Directorate of Education Administration and New Student Admission of IPB University shows that there are 5,345 schools that their graduates have registered and accepted at IPB University from second semester of 2013 until first semester of 2019. Based on Table 2 for each semester, the majority of schools included in category D, where this category indicated that there were no graduates from the schools that registered at IPB or there were graduates who registered but were not accepted as students at IPB.

The highest percentage of schools that included in the A+ category was in the second semester of 2015, which was 11.34%. As for the percentage of schools that include in the A, A-, B +, B and B- categories, there are at most respectively in the first semester of 2018, second semester of 2018, second semester of 2013, second semester of 2013, and second semester of 2017. While for the C +, C, and C- categories the amount was classified as very small and does not differ greatly between categorical variables.

Table 2: Descriptive statistics of categorical variables

Variables	Percentage of categorical data										Total
	A+	A	A-	B+	B	B-	C+	C	C-	D	
X1_2nd semester of 2013	10.34	6.87	1.87	0.49	0.67	0.15	0.02	0.09	0.04	79.46	100.00
X2_1st semester of 2014	9.78	8.08	1.91	0.28	0.36	0.17	0.00	0.04	0.02	79.36	100.00
X3_2nd semester of 2014	10.63	7.33	1.87	0.28	0.37	0.17	0.00	0.00	0.06	79.29	100.00
X4_1st semester of 2015	9.80	8.38	2.22	0.06	0.21	0.19	0.02	0.00	0.02	79.10	100.00
X5_2nd semester of 2015	11.34	7.20	1.68	0.34	0.34	0.19	0.00	0.02	0.00	78.89	100.00
X6_1st semester of 2016	10.05	8.70	1.87	0.11	0.22	0.11	0.00	0.00	0.02	78.92	100.00
X7_2nd semester of 2016	8.42	8.64	2.28	0.23	0.60	0.22	0.00	0.02	0.02	79.57	100.00
X8_1st semester of 2017	8.25	9.63	2.04	0.15	0.34	0.17	0.02	0.00	0.02	79.38	100.00
X9_2nd semester of 2017	9.56	8.55	2.90	0.30	0.47	0.24	0.07	0.02	0.06	77.83	100.00
X10_1st semester of 2018	8.89	10.16	2.67	0.09	0.15	0.11	0.02	0.04	0.02	77.85	100.00
X11_2nd semester of 2018	9.24	9.32	3.03	0.19	0.19	0.13	0.00	0.00	0.02	77.88	100.00
X12_1st semester of 2019	10.65	7.26	2.19	0.39	0.47	0.07	0.00	0.02	0.02	78.93	100.00

Table 3 showed that from second semester of 2013 until first semester of 2019, there were schools that did not have graduates who were accepted into IPB University. However, there was also school that even from second semester of 2013 until first semester of 2019 have graduates who were accepted at IPB and successfully graduated from IPB with a total of 492 students.

On the average, there were 8 graduates from every school that accepted at IPB University as students and successfully graduated from IPB from the second semester of 2013 until first semester of 2019. Based on the Grade Point Average (GPA) in IPB University, on average there are two students from each school who have relatively high GPA, which is a minimum of 3.50. Whereas there are four students on average from each school with a GPA between 2.75 and 3.50, and there are two students on average from each school with the GPA less than 2.75.

Table 3: Descriptive statistics of numerical variables

Numerical variables	Average	Minimum	Maximum	Standard deviation
X13_The number of undergraduate students	8	0	492	21.5
X14_GPA ≥ 3.50	2	0	197	7.8
X15_GPA ≥ 2.75 and < 3.50	4	0	208	10.5
X16_GPA < 2.75	2	0	136	5.4

The next step before processing data using the k-prototypes algorithm, the standardization of numerical data (z-score) is carried out. This is important so that the data becomes more uniform or there is no data with a very small or very large values.

4.2 The optimal number of cluster's selection

One of the methods of the nonhierarchical clustering is the k-prototypes algorithm. The most important step in this clustering process is determining the number of clusters.

This is important because the different number of clusters will result different characteristics of the clusters too, so the result of the conclusions will be different too. Therefore, it is necessary to find the optimal number of clusters before the algorithm was run. The optimal number of clusters is determined by the ratio between the the standard deviation within cluster (Sw) and standard deviation between cluster (Sb). The smaller value of the ratio, it can be interpreted that the diversity within the cluster is also getting smaller or the conditions to be homogeneous and the diversity between the clusters is greater or the conditions tend to be heterogeneous, so that the result of the clustering is getting better.

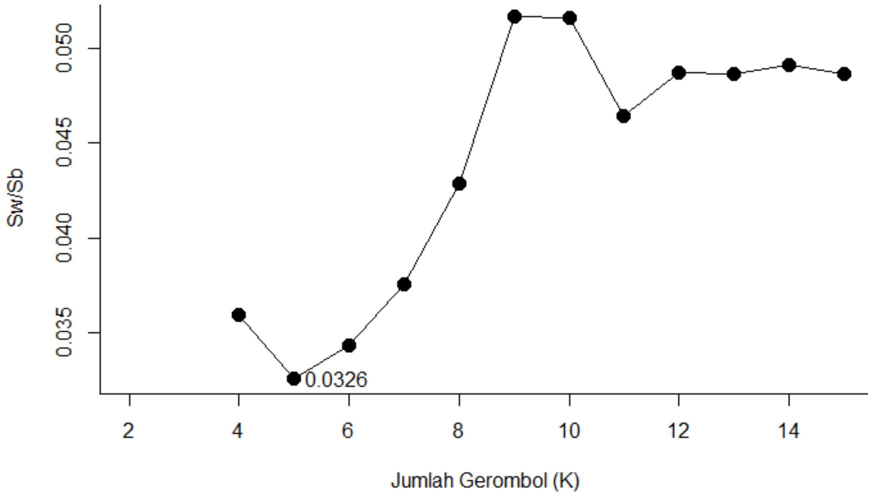

Figure 1: The ratio of Sw and Sb to determine the optimal number of clusters.

Table 4: The ratio of Sw and Sb

k	Sw	Sb	Sw/Sb
2	0.5686	1.1921E-07	Inf
3	0.5686	8.4294E-08	Inf
4	0.4645	13.6596	0.0359
5	0.4263	13.5700	0.0326
6	0.3717	13.4623	0.0343
7	0.3717	12.2893	0.0376
8	0.3943	10.9487	0.0429
9	0.4141	9.6910	0.0517
10	0.4369	8.7670	0.0516
11	0.3821	9.5066	0.0464
12	0.3821	9.0641	0.0487
13	0.3728	8.8496	0.0486
14	0.3689	8.5846	0.0491
15	0.3604	8.4227	0.0487

The cluster's distribution in the k-prototypes algorithm will be different for each reprocessing even though the processing is using the same number of clusters. It can happen because the determination of cluster center initialization in each processing is random. Therefore it is necessary to set the set.seed value in the initial processing, so that the clustering distribution values that obtained are the same for each reprocessing and in this study use the set.seed (100) value.

The selection of the number of clusters is starting from 2 to 15 clusters. Based on Table 4 it can be seen that for 2 and 3 clusters (k), the ratio value that obtained is very large, so it is not included in Figure 1. There was a decreasing value of the ratio when k = 4 to k = 5, then subsequently increased to k = 9 and decreased again until k = 11, until finally rising slowly until k = 15 (Figure 1). The smallest value of the ratio between Sw and Sb is shown when k = 5, which is equal to 0.0326. Therefore, the optimal number of clusters that will be used further in the k-prototypes algorithm is 5 clusters. The complete ratio values of Sw and Sb for each cluster size can be seen in Table 4.

The formation of clusters in this algorithm is also determined by the weighting coefficient (γ). Based on processing using R software, the weighting coefficient (γ) is obtained by 2.7724 and this value is the same for each number of clusters from 2 to 15 clusters. The value of the weighting coefficient is obtained based on the number of observations, the number of categorical variables, and the number of numerical variables.

4.3 Cluster’s member

The processing of k-prototypes algorithm using 5 clusters (k = 5) indicated the process of elimination one cluster because the last cluster did not have a cluster member, so for the end we only used 4 clusters. Based on Table 5, cluster 1 is the cluster that has the most members, as many as 4,164 schools, while the cluster with the fewest members is cluster 2. There is too many number of schools that used in this study, so all of the code of school per cluster cannot be displayed in the discussion.

Table 5: The distribution of cluster’s member

Cluster	The number of schools	Percentage
1	4,164	77.90
2	369	6.90
3	391	7.32
4	421	7.88
Total	5,345	100.00

4.4 The characteristics of cluster based on categorical variables

The characteristics of the four clusters can be seen based on each variable. The characteristics of the cluster when seen from its categorical variables are shown in Table 6 to Table 15. We can see from some of these tables that the first cluster is dominated by A- and D categories, whereas the second cluster is dominated by A, B, and C categories. Furthermore, the third cluster is dominated by B- and C- categories, while the last cluster is dominated by A +, B +, and C + categories.

Table 6: Percentage for A+ of schools's predicate

Categorical variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
X1_2nd semester of 2013	7.96	18.81	16.09	57.14	100.00
X2_1st semester of 2014	8.60	17.59	16.63	57.17	100.00
X3_2nd semester of 2014	10.39	15.14	14.08	60.39	100.00
X4_1st semester of 2015	9.54	12.98	14.69	62.79	100.00
X5_2nd semester of 2015	16.34	17.82	12.21	53.63	100.00
X6_1st semester of 2016	13.78	13.78	12.48	59.96	100.00
X7_2nd semester of 2016	13.11	13.78	10.22	62.89	100.00
X8_1st semester of 2017	12.24	13.61	9.98	64.17	100.00
X9_2nd semester of 2017	17.03	17.22	8.81	56.95	100.00
X10_1st semester of 2018	15.79	16.42	8.21	59.58	100.00
X11_2nd semester of 2018	17.41	17.81	8.50	56.28	100.00
X12_1st semester of 2019	26.36	16.34	9.14	48.15	100.00

Table 7: Percentage for A of schools's predicate

Categorical variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
X1_2nd semester of 2013	0.27	23.16	59.95	16.62	100.00
X2_1st semester of 2014	0.46	25.69	55.79	18.06	100.00
X3_2nd semester of 2014	21.17	47.70	20.66	10.46	100.00
X4_1st semester of 2015	19.87	47.54	17.86	14.73	100.00
X5_2nd semester of 2015	21.30	48.05	16.10	14.55	100.00
X6_1st semester of 2016	21.72	46.24	16.13	15.91	100.00
X7_2nd semester of 2016	21.00	46.10	10.39	22.51	100.00
X8_1st semester of 2017	22.14	45.83	10.87	21.17	100.00
X9_2nd semester of 2017	29.54	40.92	11.60	17.94	100.00
X10_1st semester of 2018	27.44	40.70	13.26	18.60	100.00
X11_2nd semester of 2018	36.75	24.70	19.68	18.88	100.00
X12_1st semester of 2019	45.62	23.20	18.30	12.89	100.00

Table 8: Percentage for A- of schools's predicate

Categorical variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
X1_2nd semester of 2013	28.00	23.00	44.00	5.00	100.00
X2_1st semester of 2014	28.43	24.51	41.18	5.88	100.00
X3_2nd semester of 2014	40.00	33.00	23.00	4.00	100.00
X4_1st semester of 2015	42.02	26.05	27.73	4.20	100.00
X5_2nd semester of 2015	48.89	26.67	15.56	8.89	100.00
X6_1st semester of 2016	50.00	30.00	13.00	7.00	100.00
X7_2nd semester of 2016	50.00	24.59	22.13	3.28	100.00
X8_1st semester of 2017	54.13	16.51	26.61	2.75	100.00
X9_2nd semester of 2017	56.13	18.71	21.94	3.23	100.00
X10_1st semester of 2018	61.54	13.29	20.98	4.20	100.00
X11_2nd semester of 2018	66.05	12.96	17.28	3.70	100.00
X12_1st semester of 2019	62.39	12.82	17.95	6.84	100.00

Table 9: Percentage for B+ of schools’s predicate

Categorical variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
X1_2nd semester of 2013	0.00	23.08	23.08	53.85	100.00
X2_1st semester of 2014	0.00	26.67	13.33	60.00	100.00
X3_2nd semester of 2014	0.00	26.67	13.33	60.00	100.00
X4_1st semester of 2015	0.00	33.33	33.33	33.33	100.00
X5_2nd semester of 2015	0.00	16.67	16.67	66.67	100.00
X6_1st semester of 2016	16.67	33.33	0.00	50.00	100.00
X7_2nd semester of 2016	16.67	25.00	0.00	58.33	100.00
X8_1st semester of 2017	12.50	25.00	12.50	50.00	100.00
X9_2nd semester of 2017	0.00	31.25	6.25	62.50	100.00
X10_1st semester of 2018	0.00	60.00	20.00	20.00	100.00
X11_2nd semester of 2018	10.00	30.00	20.00	40.00	100.00
X12_1st semester of 2019	23.81	23.81	4.76	47.62	100.00

Table 10: Percentage for B of schools’s predicate

Categorical variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
X1_2nd semester of 2013	5.56	30.56	55.56	8.33	100.00
X2_1st semester of 2014	5.26	15.79	52.63	26.32	100.00
X3_2nd semester of 2014	20.00	35.00	15.00	30.00	100.00
X4_1st semester of 2015	9.09	36.36	18.18	36.36	100.00
X5_2nd semester of 2015	16.67	50.00	11.11	22.22	100.00
X6_1st semester of 2016	8.33	75.00	8.33	8.33	100.00
X7_2nd semester of 2016	18.18	45.45	21.21	15.15	100.00
X8_1st semester of 2017	16.67	44.44	16.67	22.22	100.00
X9_2nd semester of 2017	8.00	40.00	28.00	24.00	100.00
X10_1st semester of 2018	25.00	50.00	12.50	12.50	100.00
X11_2nd semester of 2018	30.00	30.00	10.00	30.00	100.00
X12_1st semester of 2019	24.00	40.00	16.00	20.00	100.00

Table 11: Percentage for B- of schools’s predicate

Categorical variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
X1_2nd semester of 2013	0.00	50.00	50.00	0.00	100.00
X2_1st semester of 2014	22.22	33.33	44.44	0.00	100.00
X3_2nd semester of 2014	22.22	33.33	33.33	11.11	100.00
X4_1st semester of 2015	10.00	30.00	20.00	40.00	100.00
X5_2nd semester of 2015	30.00	10.00	50.00	10.00	100.00
X6_1st semester of 2016	50.00	16.67	33.33	0.00	100.00
X7_2nd semester of 2016	25.00	50.00	16.67	8.33	100.00
X8_1st semester of 2017	33.33	44.44	22.22	0.00	100.00
X9_2nd semester of 2017	46.15	46.15	7.69	0.00	100.00
X10_1st semester of 2018	66.67	33.33	0.00	0.00	100.00
X11_2nd semester of 2018	57.14	14.29	28.57	0.00	100.00
X12_1st semester of 2019	25.00	0.00	0.00	75.00	100.00

Table 12: Percentage for C+ of schools's predicate

Categorical variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
X1_2nd semester of 2013	0.00	0.00	100.00	0.00	100.00
X2_1st semester of 2014	0.00	0.00	0.00	0.00	100.00
X3_2nd semester of 2014	0.00	0.00	0.00	0.00	100.00
X4_1st semester of 2015	0.00	0.00	0.00	100.00	100.00
X5_2nd semester of 2015	0.00	0.00	0.00	0.00	100.00
X6_1st semester of 2016	0.00	0.00	0.00	0.00	100.00
X7_2nd semester of 2016	0.00	0.00	0.00	0.00	100.00
X8_1st semester of 2017	0.00	0.00	0.00	100.00	100.00
X9_2nd semester of 2017	0.00	25.00	25.00	50.00	100.00
X10_1st semester of 2018	0.00	0.00	0.00	100.00	100.00
X11_2nd semester of 2018	0.00	0.00	0.00	0.00	100.00
X12_1st semester of 2019	0.00	0.00	0.00	0.00	100.00

Table 13: Percentage for C of schools's predicate

Categorical variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
X1_2nd semester of 2013	0.00	40.00	40.00	20.00	100.00
X2_1st semester of 2014	0.00	50.00	50.00	0.00	100.00
X3_2nd semester of 2014	0.00	0.00	0.00	0.00	100.00
X4_1st semester of 2015	0.00	0.00	0.00	0.00	100.00
X5_2nd semester of 2015	0.00	0.00	0.00	100.00	100.00
X6_1st semester of 2016	0.00	0.00	0.00	0.00	100.00
X7_2nd semester of 2016	0.00	0.00	100.00	0.00	100.00
X8_1st semester of 2017	0.00	0.00	0.00	0.00	100.00
X9_2nd semester of 2017	0.00	100.00	0.00	0.00	100.00
X10_1st semester of 2018	0.00	100.00	0.00	0.00	100.00
X11_2nd semester of 2018	0.00	0.00	0.00	0.00	100.00
X12_1st semester of 2019	0.00	100.00	0.00	0.00	100.00

Table 14: Percentage for C- of schools's predicate

Categorical variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
X1_2nd semester of 2013	0.00	100.00	0.00	0.00	100.00
X2_1st semester of 2014	0.00	100.00	0.00	0.00	100.00
X3_2nd semester of 2014	0.00	0.00	33.33	66.67	100.00
X4_1st semester of 2015	0.00	0.00	100.00	0.00	100.00
X5_2nd semester of 2015	0.00	0.00	0.00	0.00	100.00
X6_1st semester of 2016	0.00	0.00	100.00	0.00	100.00
X7_2nd semester of 2016	0.00	0.00	100.00	0.00	100.00
X8_1st semester of 2017	0.00	0.00	100.00	0.00	100.00
X9_2nd semester of 2017	66.67	0.00	33.33	0.00	100.00
X10_1st semester of 2018	0.00	0.00	100.00	0.00	100.00
X11_2nd semester of 2018	0.00	0.00	0.00	100.00	100.00
X12_1st semester of 2019	100.00	0.00	0.00	0.00	100.00

Table 15: Percentage for D of schools's predicate

Categorical variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
X1_2nd semester of 2013	96.28	3.11	0.12	0.49	100.00
X2_1st semester of 2014	96.30	3.04	0.09	0.57	100.00
X3_2nd semester of 2014	93.82	1.16	4.67	0.35	100.00
X4_1st semester of 2015	93.97	1.16	4.61	0.26	100.00
X5_2nd semester of 2015	93.27	0.92	5.48	0.33	100.00
X6_1st semester of 2016	93.27	0.90	5.50	0.33	100.00
X7_2nd semester of 2016	92.57	0.94	6.09	0.40	100.00
X8_1st semester of 2017	92.62	0.97	6.01	0.40	100.00
X9_2nd semester of 2017	92.43	1.01	5.96	0.60	100.00
X10_1st semester of 2018	92.43	0.96	5.94	0.67	100.00
X11_2nd semester of 2018	90.80	3.12	5.24	0.84	100.00
X12_1st semester of 2019	88.91	3.67	5.74	1.68	100.00

If we looked from the dominance of school categories that included in each cluster, then it can be said that based on its categorical variable, the fourth cluster is the best cluster related to student admission at IPB University. It means that the fourth cluster has the best predicate or it can be also said that every school in the fourth cluster is a very good school according to the student admission at IPB University. The second and third positions were respectively occupied by the second cluster and the first cluster according to the best predicate of student admission at IPB, while the third cluster occupied the last position as the worst cluster or can it also said that every schools in this cluster was not good according to the student admission at IPB.

Table 16: The descriptive statistics of each cluster

Numerical Variables		Cluster 1	Cluster 2	Cluster 3	Cluster 4
X13_The number of undergraduate student	Average	1.26	20.02	10.09	58.29
	Standard deviation	3.66	9.89	7.84	49.91
	Total	5,263.00	7,388.00	3,946.00	24,542.00
X14_GPA ≥ 3.50	Average	0.41	3.28	2.53	20.46
	Standard deviation	1.86	2.98	4.03	18.96
	Total	1,699.00	1,209.00	990.00	8,612.00
X15_GPA ≥ 2.75 and < 3.50	Average	0.54	11.58	4.85	28.09
	Standard deviation	2.01	6.56	3.82	23.60
	Total	2,235.00	4,274.00	1,898.00	11,825.00
X16_GPA < 2.75	Average	0.40	5.16	2.71	9.75
	Standard deviation	1.18	4.83	3.33	15.25
	Total	1,670.00	1,905.00	1,058.00	4,105.00

4.5 The characteristics of cluster based on numerical variables

Based on the number of undergraduate student variables, the fourth cluster is the cluster which has the greatest number of graduates who have received and graduated from IPB University (Table 16). During the second semester of 2013 until the first semester of 2019, there were 24,542 total graduates from schools in the fourth cluster that accepted and graduated from IPB, in other words the schools that included in the fourth cluster were schools with a very good predicate related to the student admission at IPB University. The next position with the highest number of graduates was in the second and the first cluster, while the third cluster was a cluster with a less good predicate according to the student admission at IPB University. The same phenomenon was also seen in almost all other numerical variables. Based on Figure 2 to Figure 5 also strengthened the evidence that the fourth cluster was the best cluster because it has a greater distribution value than the other clusters for each numerical variable.

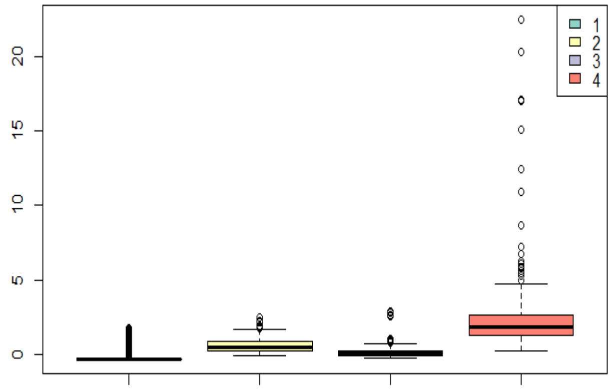

Figure 2: The characteristics of each cluster based on X13_The number of undergraduate student

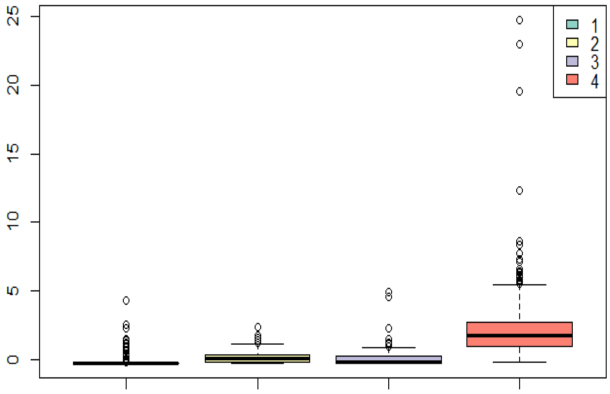

Figure 3: The characteristics of each cluster based on X14_GPA ≥ 3.50

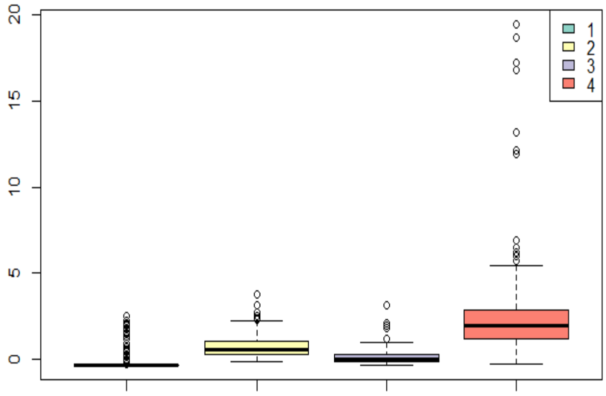

Figure 4: The characteristics of each cluster based on X15_GPA ≥ 2.75 and < 3.50

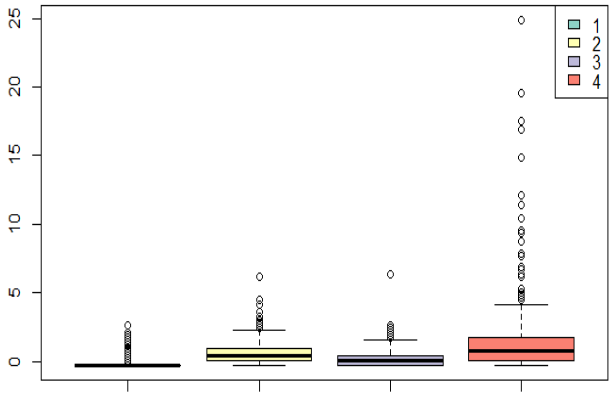

Figure 5: The characteristics of each cluster based on X16_GPA < 2.75

5. Conclusions

Based on the k-prototypes algorithm, the optimal number of clusters that used in this study were 5 clusters. However, there was a reduction in the number of clusters because the last cluster did not have any cluster members, so the final result was 4 clusters. The first cluster has 4,164 members, the second cluster has 369 members, the third cluster has 391 members, while the last cluster has 421 members. If seen based on its categorical variables, the fourth cluster was the best cluster related to student admission at IPB University. The fourth cluster has the best predicate cluster or it can be also said that every school in this cluster was a very good school according to the student admission at IPB University. The second, third, and last position with the best predicate school are respectively occupied by the second, the first, and the third cluster. If seen based on its numerical variables, the fourth cluster was a cluster which has the most number of graduates who have received and graduated from IPB University, so every schools in the fourth cluster has a very good predicate related to student admission at IPB University. The second, third, and last position with the highest number of graduates are respectively occupied by the second, the first, and the third cluster.

The characteristics of the cluster looked the same, both based on categorical and numerical variables, where the fourth cluster was the best cluster, followed by the second and the first cluster. While the third cluster was the worst cluster according to the student admission at IPB University.

References

Anderberg, M. R. (1973). *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks* (Vol. 19). New York (US): Academic press.

Bunkers, M. J., Miller, J. R., & DeGaetano, A. T. (1996). Definition of climate regions in the Northern Plains using an objective cluster modification technique. *Journal of Climate*, *9*(1): 130-146.

Huang, Z. (1998). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining, (PAKDD)* (pp. 21-34).

Kader, G. D., & Perry, M. (2007). Variability for categorical variables. *Journal of Statistics Education*, *15*(2). DOI: 10.1080/10691898.2007.11889465.

Li, Y., Luo, C., & Chung, S. M. (2008). Text clustering with feature selection by using statistical data. *IEEE Transactions on knowledge and Data Engineering*, *20*(5): 641-652.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, *38*(11): 1857-1874.

Lin, H. J., Yang, F. W., & Kao, Y. T. (2005). An efficient GA-based clustering technique. *Journal of Applied Science and Engineering*, *8*(2): 113-122.

Okada, T. (1999). Sum of squares decomposition for categorical data. *Kwansei Gakuin Studies in Computer Science*, *14*(1): 1-6.

Rencher, A. C. (2007). *Methods of multivariate analysis* [second edition]. John Wiley & Sons.