ISSN 2278-3091

Volume 9, No.1.3, 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse5791.32020.pdf

https://doi.org/10.30534/ijatcse/2020/5791.32020

Analyzing the College Freshmen Aptitude Results using K-means Algorithm

Lourdes M. Padirayon[1], Manny S. Alipio[2], Freddie P. Masuli[3], Grecilia A. Callitong[4], Daniel T. Ursulum[5]

[1] Cagayan State University-Sanchez Mira, Philippines, desmpadirayon@gmail.com
[2] Cagayan State University-Sanchez Mira, Philippines, msa_859@yahoo.com
[3] Cagayan State University-Sanchez Mira, Philippines, salacnib@yahoo.com
[4] Cagayan State University-Sanchez Mira, Philippines, callitonggrace@yahoo.com.ph
[5] Cagayan State University-Sanchez Mira, Philippines, ursulumjr_daniel@yahoo.com

ABSTRACT

Universities today are operating in a very complex and highly competitive environment. Huge data available in the educational field bring out the hidden knowledge which is helpful in decision making.

Cagayan State University used Rapidminer as an educational data mining (EDM) tool to extract new knowledge from huge standardized aptitude result of 3,393 data stored in an Excel database. Clustering algorithms such as K-Means where k=3 and K-means where k=4 number were used for evaluating students' records run over the Rapidminer application to enable clustering prototype.

The prototype was assessed and used internal validation using the silhouette index to classify which clustering algorithms work best. Finally, the study determined the degree of the students' aptitude that can be used for advising students on the most appropriate course that they can take. The K-means algorithms using parameter k=4 was determined as the best parameter in the clustering.

Key words: Data mining, Clustering algorithm, K-means, Educational Data Mining, Student's Aptitude

1. INTRODUCTION

Today, universities work in a very dynamic and highly competitive world [1]. The Cagayan State University-Sanchez Mira Campus is not an exemption. Thousands of students from different secondary high schools of nearby towns and provinces flocked to the University for the College Entrance Examination in order to have a chance in enrolling in the university. Along this, the university is always trying to improve the college admission examination system. The accuracy of course placement would improve learning outcomes [2] and taking proper intervention at correct time can increase the quality of learning.

Furthermore, in assessing educational institutions, educational data mining is an approach to explore the specific types of data from educational backgrounds and this technique considers the knowledge of students acquired [3]. The Cagayan State

University-Sanchez Mira Campus used EDM educational data mining to extract new knowledge from huge standardized aptitude results which are stored and increasing rapidly. Huge data available in the educational field can give the hidden knowledge from it.

Data Mining is an effective technique used to derive any useful information from a large database by analyzing data from different perspectives [4] and to find data segmentation and pattern information. By clustering algorithm, the process can give the characteristics of each observed cluster, and make further analysis and interpretation on the whole data set. [5]

Clustering is a concise work attempting to identify related clusters of artifacts formed from the properties of the criteria. Present methods of clustering may be largely classified into three categories: partitioned, classified and locality-based processes [3]. The process creates a division of the different entities into clusters from which to determine the amount to be reduced [5].

Once the algorithm is running and the clusters are defined, any particular data can be assigned easily to the same category [3]. This method helps identify the students who need special advising and counseling by the concerned teacher to improve delivery of learning along the thrust of quality education.

In addition, a number of algorithms are available for analyzing students' data. Some of these are DBSCAN, K-Medoids and K-Means. K-means is one of the simplest and most collective unsupervised device-learning algorithms [7]. It is effectively used to segment a defined dataset into k groups, where k represents the number of groups or clusters [8]. The number of clusters presumes an appropriate data classification and segregates the data into clusters in such a way that the values of the data within the same cluster are identical [9]. In the clustering of k-means, each group is characterized by its center point or mean, which is also called the centroid and is defined as the mean data value within that cluster.

Enormous data can be easily be clustered using K-means clustering algorithms. One of its key benefits is the minimization of the period necessary to process. It is used for mixed datasets which is appropriate for information that are usually process in the educational field. The educational institution can gain valuable potential information by grouping the student data [10] for clustering facilitate large data sets to discover significant groupings of objects and established separate sample categories [11]. Also, the clustering method provides useful additional knowledge for exploring the creation of innovative ideas and has significant importance in imaginative education. [12]

The Cagayan State University-Sanchez Mira Campus is introducing new techniques and strategies to improve and develop the skills of its students. The university is practicing open admission to all students who were able to pass the College Freshmen Aptitude Test.

The Cagayan State University (CSU) Sanchez Mira Campus is composed of seven (7) colleges namely: Information Technology, Teacher Education, Agriculture, Industrial Technology, Arts and Sciences, Hospitality Industry Management, and Criminal Justice Administration wherein the total population is about 3,567 students in S.Y. 2015-2017. Students of the university came to from the nearby towns and provinces. As a university, CSU-Sanchez Mira prepares students to be well-educated, highly-skilled and refine-mannered according to the needs and requirements of the dynamically growing market.

The key aim of College Freshmen Aptitude Test is to assess the student's cumulative data in various ability areas such as auditory, numerical, logical and script [13]. These tests are not designed to measure what the student learned in school; rather, they measure the potential of the student-applicant in performing well in the future after college. From the college aptitude result, the student-applicant is advised to take a corresponding course that seemed to be suited according to the perceived strength and identified weaknesses. Thus, this study was conducted.

The key goal of this research is to group students based on the results of their CFAT in order to give appropriate advice and counsel to students for their most suited corresponding course that could lead the student in finishing the course on-time, lessen the case of dropped-outs, shiftees and student transferring of students to other schools.

To be able to realize the grouping of students according to their aptitude level, clustering algorithm with RapidMiner technology was utilized in this study. The dataset was verified with an internal testing method to select which among K-means clustering parameters is the most appropriate. The clustering model can produce a basic resource to promote proper admission according to student's aptitude level in the Sanchez Mira Campus of the Cagayan State University.

1.1. Conceptual Framework

The context of the study was established on the Knowledge Discovery Process (KDP) by S.H. Moad . et.al [13] as shown in figure 1.

1. Selection; the source of the data is described in.
2. Preprocessing; seeks dataset changes.
3. Conversion; it analyzes factors, their meaning, their relationship and their correlation.
4. Datamining; this involves the usage of mathematical learning approaches to extract data, correlations, parameters and rules.
5. Interpretation / Appraisal; shows the weather patterns observed which can be either stimulating or not.

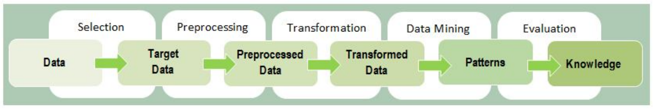


Figure 1: Knowledge Discovery Process

The KDP statistic has been modified to suit with the study's objectives. Data extraction is only a part of the entire research context. Figure 2 indicates the structure of the clustering method as its core modules included in this analysis. The data stored in the guidance counselor's office were examined to be able to collect the correct dataset for the analysis. After pre-processing the dataset was developed. This functioned influence the data mining instrument for the application of the particular K-means=3 and K-means=4 algorithms to be able to realize the group of students that need course reassignment. Four clusters were created after the procedure. The cluster models were tested with the findings obtained from the dataset focused on their correctness. Cluster quantity reflects classes that are connected to the student or to each other. The discovered information could instead be used for decision-making.

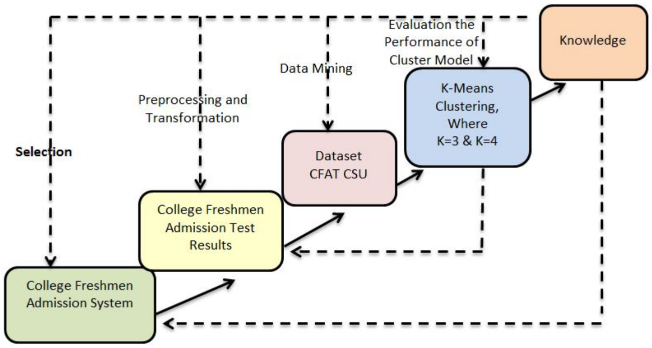

Knowledge
Evaluation the Performance of Cluster Model
Data Mining
K-Means Clustering, Where K=3 & K=4
Preprocessing and Transformation
Dataset CFAT CSU
Selection
College Freshmen Admission Test Results
College Freshmen Admission System

Figure 2: Steps of mining information from data.

1.2. The Research Aims

The main purpose of this study is to discover the College Freshmen Aptitude Test data set in clustering for identifying the levels of aptitude of students-applicants

Specifically, the College Freshmen Admission Test records were used to realize the following objectives:

1. to cluster students based on the result of their College Freshmen Aptitude Test in order to be used in advising and counseling students for their most appropriate course.
2. to identify the best Clustering algorithm in processing the data set
3. to evaluate data using the best clustering algorithm and tailor-fitting using different parameters

2. METHODOLOGY

2.1. Data Set

The data mining instrument used in this study is RapidMiner which deals diverse data mining methods for numerous types of data. It introduces the utmost significant machine learning algorithms, data preprocessing, and transformation methods. Varied approaches are available in data science to group the data and build based on closeness among the information, density in the dataset, or new neural network presentation. Any of them are K-Means forms of clustering, intensity clustering and self-organization maps [15].

The dataset used in this study was collected through writing a request letter from the head of the campus, asking permission to collect the needed data for the study. The initial size of the dataset is 3,393 records.

Data collected were the results of the college freshmen aptitude tests of the Cagayan State University-Sanchez Mira for the years 2015-2017. These College Freshmen Aptitude Test results were compiled in the office of the guidance counselor. See Table 1, the sample data set.

Table 1: Sample CFAT Data

Attribute	Values	Description
Schools Last Attended	1-58	Flora National High School
Mechanical Reasoning	6 above 4-5 3 below	Above Average (AA) Average (A) Below Average (BA)
Language Usage	4-5 3 below 6 above	Above Average (AA) Average (A) Below Average (BA)
Numerical Reasoning	4-5 3 below 6 above	Above Average (AA) Average (A) Below Average (BA)
Abstract Reasoning	4-5 3 below 6 above	Above Average (AA) Average (A) Below Average (BA)
Course	1-7	BEED, BS Info Tech, BSA, BSACT, BS Crim, BSE, BHIM, BSIT, Psych, DCA

The data collected were entered into the Microsoft Excel Application for data cleansing or attributes needed to be transformed into nominal data. Standardizing the data involved the following steps : 1. removing extra spaces, 2. filling all blank cells with '0', 3. translating numbers kept as text into quantities, 4. eliminating duplicate values from the data set, 5. altering text to lower or upper case for consistency, and 6. checking the spelling.

The standardized data set is transformed by saving the data into csv (comma separated values) format. See table 2. After standardizing or data cleansing, these were entered into the Rapid Miner application for pre-processing. Finally, data analysis is performed to find out that test result's attributes that served basis for clustering on the various scopes of academic abilities of the student's possess.

Table 2: College Freshmen Aptitude Test Results

School	Mechanical Reasoning	Language Usage	Numerical Reasoning	Abstract Reasoning	Priority Course
1	2	2	2	3	BSA
1	2	1	3	2	BSIT
2	4	2	3	7	BEED
2	1	3	4	4	BSA
2	1	2	1	2	BSA
2	2	1	3	4	Accounting Technology
2	2	3	5	4	Agricultural Engineering
2	1	2	2	2	Agricultural Engineering

2.2. Clustering

Clustering is a method of categorizing the raw facts and examines the unseen patterns that may occur in datasets. It is a procedure of grouping data objects into fragmented clusters so that the data in the same cluster are comparable, yet data fitting to dissimilar cluster change. Among the techniques of data mining, clustering supports to categorize the group of students according to CFAT results.

In this study, the clustering algorithms are assessed using internal validity methods to select which of the algorithms is utmost appropriate for the CFAT dataset. One clustering algorithm was exposed to the internal assessment which is the k-means.

K-means clustering method was chosen to analyze the data set extracted from the results of the College Freshmen admission test. A clustering algorithm tries to discover clusters of data based on some comparison. Also, the clustering algorithm determines the centroid of a group of data sets [16]. Some algorithms measure the gap between a point and the cluster centroids to control cluster association. The product of a clustering algorithm is a statistical analysis of the cluster centroids with the number of components in-cluster. Following, the example is allocated to the near centroid by means of a metric, usually recognized as the Euclidean distance. Later, the centroid is recalculated using the new group formed. This procedure lasts until a criterion is achieved. [17]

2.3. Internal Validation Measure

The cluster examination contains of the clustering outcome assessment, in order to find the partition that well matches the data [18]. Inner validation is the easiest way to test a clustering algorithm if it is useful to a data collection, from the time it uses; only the distinct scattering of the points and the cluster labels generated by the algorithm measure the cluster properties.

This collection of approaches is focused on the premise that the algorithms will be investigated for clusters whose association are near to each other and far from other cluster members. Also, establish number of comparisons which are the source of the approximation algorithm's effectiveness [19]. In this analysis, it measured the typical silhouette width of k-means=3 and k-means=4, the maximization of insights. The computational silhouette index I method is as follows [20]:

For a detailed cluster, Xj (j = 1, .. c), the procedure of silhouette allocates a reliability metric to the Xj ith array, s(i)=(i = 1, ... m), defined as the width of the silhouette. This value is an indicator of trust about the sample membership in the Xj cluster, and is defined as:

$$s(i) = \frac{(b(i) - a(i))}{Max\{a(i), b(i)\}}$$

Where a(i) is the maximum distance between the ith sample and all samples involved in the Xj; b(i) is the total minimum distance between the sh sample and all samples included in Xk(k = 1,. c; k j).

Silhouette was first defined by Rousseau P.J. 1987 [20]. The silhouette meaning is a function of how similar an object is to its own cluster when contrasted with other clusters. Often, it ranges from −1 to +1, where the node is closely linked to its own cluster and not well matched with adjacent clusters with a large interest. When the plurality of artifacts has a high value otherwise the clustering structure is correct [18].

2.4. Clustering using RapidMiner

The clustering and assessment method for the clustering algorithms is revealed in Fig. 3.

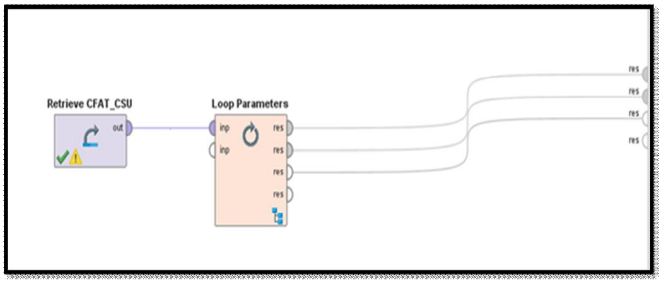


Figure 3: Rapidminer Procedure for Clustering and Evaluating Clustering Algorithm

The Rapidminer operator accessed the CFAT data from the source where the CFAT dataset was stored. The RapidMiner loaded the data and converted into complete metadata which helped as the output of the operator.

In addition, the 'Loop Parameters' operator was specified to conduct a loop evaluation method of the numerous algorithm parameters used. The ‘Loop Parameters’ operator covers a ‘Clustering (Subprocess)’ and ‘Silhouette Index’ as inner operators as shown in Fig. 4 and its formation was set over the edit bound settings.

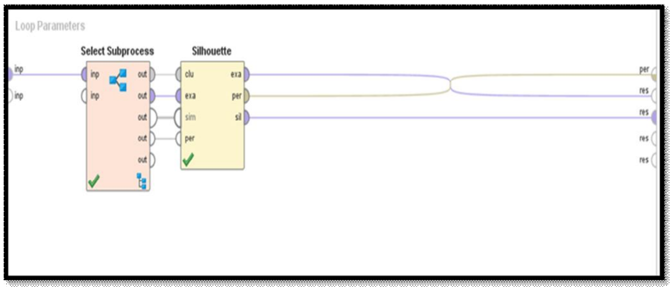
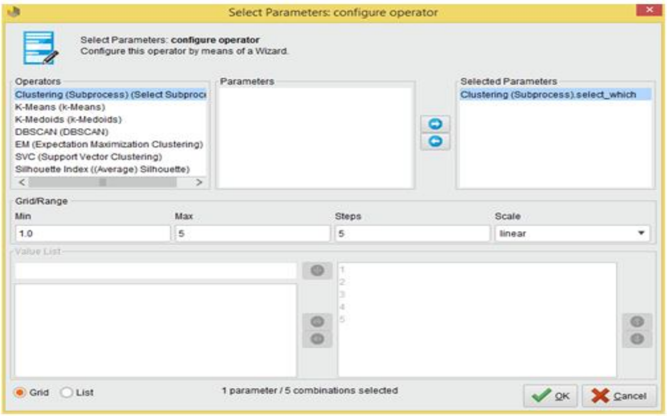


Figure 4: The inner operators of the ‘Loop Parameters’ operator

The ‘Loop Parameters’ uses the ‘Select which’ parameter of the ‘Clustering (Subprocess)’ which allows the algorithm to be performed in every repetition. Parameter series 1 to 3 with 5 periods and a linear scale was used because there were e 3 parameters evaluated. As shown in Figure 5.



Select Parameters: configure operator

Select Parameters: **configure operator**
Configure this operator by means of a Wizard.

Operators
- Clustering (Subprocess) (Select Subproc
- K-Means (k-Means)
- K-Medoids (k-Medoids)
- DBSCAN (DBSCAN)
- EM (Expectation Maximization Clustering)
- SVC (Support Vector Clustering)
- Silhouette Index ((Average) Silhouette)

Parameters

Selected Parameters
- Clustering (Subprocess).select_which

Grid/Range

Min	Max	Steps	Scale
1.0	5	5	linear

Value List

1
2
3
4
5

● Grid ○ List

1 parameter / 5 combinations selected

Figure 5: Select parameters to configure operator

Two (2) parameters were used for internal operators of the 'Clustering (Subprocess)' such as K-means where k=3 and K-means if k = 4, as shown in Fig. 6

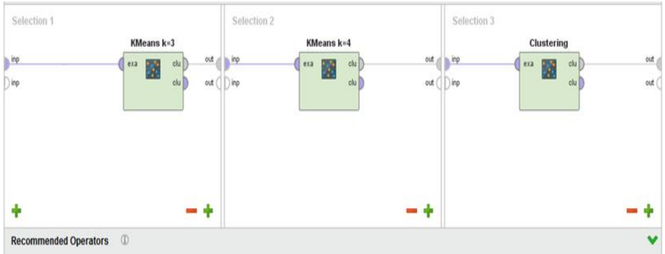


Figure 6: 'Selection' Parameter of the 'Clustering (Subprocess)'

One of these algorithms is executed in each loop of the "Vector Parameters" and its model is sent to the "Silhouette Table" operator such that all the algorithms are evaluated with their default parameters.

3. RESULTS AND DISCUSSION

The clustering algorithms developed clustering models and mean silhouette values after the method was executed as shown in Table 3.

Table 3: The result of the process (cluster count performance)

Clustering Algorithm	Clustering Model				Silhouette Output
	Cluster 0	**Cluster 1**	**Cluster 2**	Cluster 3	
Means (k = 3)	1552	1325	495	0	0.533
Means (k = 4)	960	473	781	1158	0.597

The K-means algorithm where k = 3 has a three-cluster profile index of 0.533. K=4 generated four clusters with an index of 0.597 for silhouette. Calculated index is similar to 0 showing that some of the samples are very near to the judgment border of two adjacent clusters.Nevertheless, as the greater the silhouette index value indicates a stronger clustering parameter, the researchers concluded that K-means=4 was the strongest one in contrast to the other clustering parameter examined.

Finally, the centroid plot generated by the algorithm k-means, in which k = 4 is seen in Fig. 7. The clustering findings are verified using the centroid map.

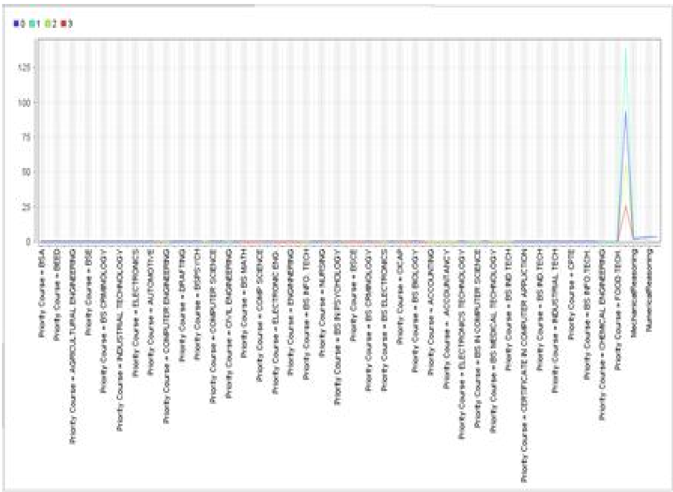


Figure76: K-means=4 Clustering Centroid Plot

From the figure it can be inferred that the students belonging to Cluster 0 (dark blue line) with 960 members, or (29 percent) of all takers, had very strong results in all the skills evaluated relative to the other three clusters. This category is labeled AVERAGE HIGHER.

Students who belong to cluster 1 (light blue), 473 (14%) takers, had the best performance in all the aptitudes compared to other clusters. They scored best on Mechanical Reasoning. Their group may be called the 'ABOVE HIGHER AVERAGE'.
Students who belong to cluster 2 (yellow) 781(23%) had to mediate performance in all the abilities tested compared to other clusters. Their group called the 'AVERAGE'.

On the other hand, cluster 3 (red), 1,158 (34%) signifies the students who had fair performance as related to the other clusters. These students achieved the poorest result. Their cluster may be called as 'BELOW BELOW AVERAGE'.

With this results, the 1,158 fair students should be given more attention in the advicing and counceling process to lead them to the most appropriate course for them.

However, all other students must still undergo a not so rigid advising and counseling as a goal of the study.

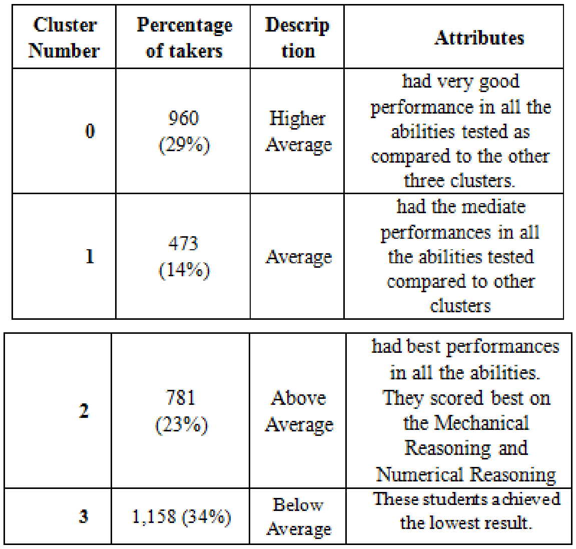
This study centered on educational data mining, utilizing the Rapidminer platform to apply a clustering algorithm on the CFAT dataset. The findings indicate that the K-means method, using k=4, given the strongest clustering model for cluster students with a value of 0.597, by applying an internal validation test which is the silhouette index test.

Four clusters were formed having 960 students in the HIGHER AVERAGE (cluster 0), 473 AVERAGE students (cluster 1), 781 ABOVE HIGHER AVERAGE students (cluster 2) and 1,158 BELOW AVERAGE students (cluster 3).

The clustering model formed by the K-means clustering algorithm can be used by school administration in promoting students collaborative learning, effective group training and personalized learning system at Cagayan State University, Sanchez Mira Campus.

As a result of the study, the researchers recommend that Cagayan State University- Sanchez Mira Campus should adopt and implement a computerized system for College Freshman. See Table 4 for description of entrants based on the clustering result

Table 4: Description of Entrants Based on the Clustering Result



Cluster Number	Percentage of takers	Description	Attributes
0	960 (29%)	Higher Average	had very good performance in all the abilities tested as compared to the other three clusters.
1	473 (14%)	Average	had the mediate performances in all the abilities tested compared to other clusters
2	781 (23%)	Above Average	had best performances in all the abilities. They scored best on the Mechanical Reasoning and Numerical Reasoning
3	1,158 (34%)	Below Average	These students achieved the lowest result.

4. CONCLUSIONS

This research centered on educational data mining, utilizing the Rapidminer method to add a clustering algorithm to the CFAT dataset. The findings indicate that the K-means method, using k=4, given the strongest clustering model for cluster students with a value of 0.597, by applying an internal validation test which is the silhouette index test.

Four clusters were created having 960 students in the HIGHER AVERAGE (cluster 0), 473 AVERAGE students (cluster 1), 781 ABOVE HIGHER AVERAGE students (cluster 2) and 1,158 BELOW AVERAGE students (cluster 3).

The clustering model produced by the K-means clustering algorithm can be used by school administration in encouraging students collaborative learning, effective group training and modified learning system at Cagayan State University, Sanchez Mira Campus.

As a result of the study, the researchers recommend that Cagayan State University- Sanchez Mira Campus should adopt and implement a computerized system for College Freshman Aptitude Test (CFAT) to ensure that a more credible advice can be given to students as to what course is well-suited for them as reflected by the result of the examination.

ACKNOWLEDGEMENT

The authors extend their gratefulness to the Cagayan State University Sanchez Mira Campus for supporting the research, providing all required resources.

REFERENCES

1. S.R.M Campos, R. Henriques, M. H. Yanaze, **Knowledge discovery through higher education census data**, Technological Forecasting and Social Change, Volume 149, 2019,
2. J. Z. Gamboa Jr. **Clustering Scholarship Programs Using Educational Data Mining Techniques**. International Journal of Advanced Trends in Computer Science and Engineering. 2019 https://doi.org/10.30534/ijatcse/2019/51832019
3. A. Hussein & A. Oluwaseun. **Data Mining Application Using Clustering Techniques (K-Means Algorithm) In The Analysis Of Student's Result.** https://www.researchgate.net/publication/333508765201 9
4. N.M. Alotaibi1.et.al. **Agent-based Big Data Mining**. International Journal of Advanced Trends in Computer Science and Engineering. 2019 https://doi.org/10.30534/ijatcse/2019/4481.12019
5. I. S. Makki and F. Alqurashi . **An Adaptive Model for Knowledge Mining in Databases "EMO_MINE" for Tweets Emotions Classification .** International Journal of Advanced Trends in Computer Science and Engineering**.** https://doi.org/10.30534/ijatcse/2018/04732018, 2018
6. I. Singh, A. S. Sabitha and A. Bansal, "**Student performance analysis using clustering algorithm,"** 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, 2016, pp. 294-299, doi: 10.1109/CONFLUENCE.2016.7508131.
7. A. Bansal and M. Sharma. **Improved K-means Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining**. International Journal of Computer Applications (0975 – 8887) Volume 157 – No 6, January 2017.
8. D. Aggarwal and D. Sharma. **Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm**. Intelligent Information International Journal of Computing Academic Research (IJCAR) ISSN 2305-9184, Volume 5, pp.88-103, 2016. DOI: 10.5120/ijca2017912719

9. K. Sya'iyah et.al. **Clustering Student Data Based On K-Means Algorithms** . (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.3, No. 8, 2012 146, 2019
10. C. Sreedhar. **Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop**.ttps://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0087-2, 2017
11. Yu-Cheng Chien, Ming-Chi Liu, Ting-Ting Wu, **Discussion-record-based prediction model for creativity education using clustering methods**, Thinking Skills and Creativity, Volume 36, 2020
12. A. M. Niessen et.al. **Admission testing for higher education: A multi-cohort study on the validity of high-fidelity curriculum-sampling tests.** PLOS ONE. https://doi.org/10.1371/journal.pone.0198746, 2018
13. S.H. Moad et.al. **Assessing reliability of Big Data Knowledge Discovery process**. (ICDS 2018).10.1016/j.procs.2019.01.005, 2019.
14. Vijay Kotu, Bala Deshpande, **Chapter 7 - Clustering, Data Science (Second Edition)**, Morgan Kaufmann, Pages 221-261, 2019
15. M. Kalra et. al. **K- Mean Clustering Algorithm Approach for Data Mining of Heterogeneous Data**. https://www.researchgate.net/publication/320932435, 2019
16. S. Soheily-Khah. **Generalized k-means based clustering for temporal data under time warp.** Artificial Intelligence [cs.AI]. Universite Grenoble Alpes, 2016. HAL Id: tel-01394280 https://hal.archives-ouvertes.fr/tel-01394280v2
17. M. Garbade. **Study of Clustering of Data Base in Education Sector Using Data Mining**. Vol. 3, Issue 10, 2015 | ISSN (online): 2321-0613, 2018
18. C. Subbalakshmia et. al. 2015. **A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set.** (http://creativecommons.org/licenses/by-nc-nd/4.0/).2015
19. R.M. Dellosa. An Efficient Position Estimation of Indoor Positioning System Based on. **International Journal of Advanced Trends in Computer Science and Engineering Dynamic Time Warping,** 2020 http://www.warse.org/IJATCSE/static/pdf/file/ijatcse0491.22020. pdfhttps://doi.org/10.30534/ijatcse/2020/0491.22020
20. P.J. Rousseeuw**. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics20**:53–65. doi:10.1016/0377-0427(87)90125-7, 1978