Normalization methods analysis of career pattern using self-organizing map

Purwanto[1], *Hapsari Peni Agustin* Tjahyaningtijas[2*], *Jesse R* Paragas[3] and *I Gusti Putu Asto* Buditjahjanto[4]

[1]Post Graduate of Electrical Engineering Department, State University of Surabaya, Indonesia.
[2]Electrical Engineering Department, State University of Surabaya, Indonesia.
[3]Information Technology Department, College of Engineering, Eastern Visayas State University, Philippine
[4]Electrical Engineering Department, State University of Surabaya, Indonesia.

Abstract. Clustering the distribution of student graduates is an approach used to analyze and understand the success of Vocational High School education programs in preparing graduates to enter the workforce or start their own businesses. The purpose of clustering is to evaluate the effectiveness of educational programs, identify entrepreneurial potential, formulate career planning, and develop entrepreneurial skills, all contributing to the fulfilment of Sustainable Development Goals (SDGs) related to quality education (SDG 4), decent work and economic growth (SDG 8), and industry, innovation, and infrastructure (SDG 9). Through this clustering, schools can evaluate the extent to which quality high school graduates achieve career or entrepreneurial success, supporting the objectives of SDG 4. This information helps in designing educational programs that are more in line with the needs of the job market, providing better career guidance to students, and promoting entrepreneurial skills among high school students, contributing to SDG 4 and SDG 8. Clustering the distribution of vocational high school students by working, continuing, and entrepreneurial status plays an important role in strengthening the link between education and the world of work, aligning with the aims of SDG 4 and SDG 8. Self-Organizing Map (SOM), as an Artificial Neural Network, assists in data clustering or mapping tasks, aiding in the discovery of patterns and trends within the vocational high school graduate population. The result of clustering using Z-Score and Min-Max normalization techniques is 5.31% and 3.98%, respectively, providing insights into the career and entrepreneurship trends and patterns of vocational high school students. This valuable information can be used for the development of educational programs, career guidance initiatives, and improved alignment between education and the needs of the world of work, ultimately contributing to the realization of SDGs 4, 8, and 9.

* Corresponding author: hapsaripeni@unesa.ac.id

1 Introduction

After completing the basic education level, students can continue vocational education at a secondary school called Vocational High School. Vocational High Schools have a more specific focus on practical learning and skills that are directly related to the world of work or industry. The effective transition of vocational high school graduates into the world of work or entrepreneurship is a crucial component of their educational journey. Distribution of graduates by graduate status can provide valuable insights for schools, policymakers, and stakeholders to improve the quality of education and enhance the alignment between education programs and labor market needs. In this context, the use of approaches to cluster and analyze the distribution of high school graduates, to address this issue can use clustering technology as a support for the work of school staff in charge of clustering the distribution of graduates followed by the students. An example of an unsupervised learning technique that creates a low-dimensional representation of the input making use of a framework for an artificial neural network is self-organizing maps [1]. Although a Self-Organizing Map is a kind of simulated artificial neural network, the way it works is slight.

Different from ordinary artificial neural networks. When an error in the Artificial Neural Network system, it performs the correction learning step using the gradient decrease. A self-organizing Map uses competitive learning, each node will compete to be the winner. A Self-organizing Map will maintain the input topological structure by applying non-stop functions.

Clustering is carried out on data on the distribution of 2019-2020 graduates and the results are displayed in the form of topological visualization of self-organizing maps. Previous research on clustering alumni data has been carried out, including research on clustering alumni data of Semarang State Polytechnic using the K-Means method and the results are presented in the form of a web-based digital map. The clustering mechanism is based on four attributes, namely: type of company, office classification, field of work, and study program competency. The test results show that 51 alumni work in accordance with their competence, while 23 alumni with less suitable jobs and 26 alumni are not in accordance with their competence. The results of this study are used to help make decisions whether to make curriculum changes or not [2]. Further research uses fuzzy C-Means to cluster graduates using GPA and length of study variables. Based on clustering, four clusters were obtained. Of the four clusters, it is known that cluster 4 has graduate members with an age range of 5.91 years. This shows that there are still many students majoring in mathematics who spend more than 10 semesters or 5 years studying. The results of this study are expected to be used as a consideration for the department in increasing the cumulative grade point average of students so that they can complete their study time quickly [3]. The Bayesian Network model was used for Hepatitis Diagnosis in [5] and combined with Relief feature selection in [6]. In [7-9], artificial neural network models were used to predict student academic performance from initial semester input grades, and in [10] combined with Linear Regression and Support Vector Regression. Self-organizing map applications have been used in unsupervised learning to group students into scholarship programs [11-12]. The Self-Organizing Map method is also used in education to show students [13-14] cognitive structure models and to track learning

activities [15]. Many studies using the SOM method have been carried out, such as in previous research named Self-Organizing Map (SOM) Algorithm-Based Applicant Mapping of Predicting Seriousness in Selecting Private University It was found that the SOM method for predicting the interest of private university registrants in 2017 was 0.0065%, in 2018 it was 0.00067%, and in 2019 it was 0.0047% [16], the next research entitled Student modelling using SOM cluster principal component analysis showed that the SOM method for modeling the learning system used the technique of clustering of the student data set using principal, the results were a map of clustering [17] , the next research is entitled the data from the student higher education census: a source of knowledge It was found that the SOM method was used to visualize profiles based on genre and academic status [18], the next research was entitled Using Self-Organizing Maps and Neural Networks, an analysis of student behavior in online learning environments using users clustering to analyze face-to-face systems, online and mixed learning systems obtained face-to-face and online results connected to a methodological approach [19].

This research will do clustering with attributes of the field of work, and cumulative grade point average. The method that will be used in clustering student distribution data is Self-Organizing Map, which is an unsupervised algorithm using Z-Score and Min-Max normalization techniques. Self-Organizing Map is chosen because in general, data cannot be explicitly separated into groups but has a tendency expressed by a degree of membership that is valued between 0 and 1 to its prevalence. This study's outcomes are anticipated to categorize the distribution of student graduates at vocational high school institutions.

2 Material and method

2.1 Self-Organizing Map (SOM)

Teuvo Kohonen initially presented SOM in 1996. Self-organizing neural networks are used by SOM to map high-dimensional data into a single lower dimension, to lower the dimensions of the data in order to view the element's target input without guidance or unsupervised learning data with the assumption of structured topology into clusters. The SOM algorithm treats each cluster unit's weight vector as an illustration of the cluster-related input pattern. Unit clusters that match the closest input vector pattern in terms of weight are considered winning. In order to match the input other winning unit and its surrounding unit keep updating their weights. Unlike previous JST models, the target neurons in the SOM network are arranged in two dimensions, with tunable shapes, as opposed to being arranged horizontally. Different forms will produce neurons surrounding different neurons that win, hence the ultimate weight may also differ. Weight adjustments are made in SOM for both the line weight to the neighboring neuron and the line weight linked to the winning neuron. Basic SOM implementations come in a variety of forms, from square long/six square, 2D/3D maps, to different distance function possibilities for defining surrounding surroundings. Implementation architecture is the implementation

SOM method, data from a high-dimensional vector space (input) each input layer neuron's connection to all other neurons in the output layer is translated into a two-dimensional vector region output at a local location. A class (cluster) of the input is represented by each neuron in the output layer. With its SOM dimensional reduction properties, it is widely used as a dimension reduction tool, such as the principal component analysis (PCA). SOM is seen as a grouping approach, but, since dimensional reduction may also be understood as a decrease in the number of dimensions (or clusters) of the data input. The formula is as follows:

$$d(i,j)=\sqrt{[\sum(x(i)-w(j))^2]} \quad (1)$$

Where:

- d(i,j) is the distance between the input vector i and the weight of the neuron j
- x(i) input vector of i.
- w(j) weight of j neuron.
- ∑ a symbol for counting.

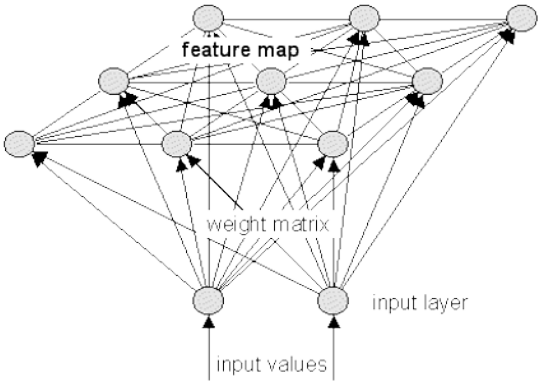

Fig. 1. Architecture Self-Organizing Map

At the training stage, each input vector will be compared to all neuron weights in the SOM. The neuron with the weight closest to the input vector will be selected as the winner (winner neuron) then the weight of the winning neuron and its surroundings will be updated (adjusted) using a calculation formula called the rule of Self-Organizing Map. These are the guidelines for the self-organizing map:

$$\Delta w(i,j) = \eta * h(i,j) * (x(i) - w(j))] \quad (2)$$

Where:

- Δw (i,j) is the change in the weight of the neuron j for the input vector i.
- η (eta) is the learning rate, which controls how much weight adjustments are made.
- h(i,j) is a responsibility function, which measures how close the j neuron is to the winning neuron.
- x(i) input vector of i.
- w(j) weight of j neuron

2.2 Normalization techniques

A data mining technique called data normalization is used to convert a dataset's values into a common scale. It is crucial to note that numerous machine learning algorithms exhibit sensitivity to the input feature scale and can yield superior outcomes upon normalization of the data. In data mining, a variety of normalization approaches can be applied, like as:

2.2.1 Normalization of min-max:

This method reduces a feature's values to a range of 0 to 1. To do this, deduct the feature's minimum value from each value, then divide the result by the feature's range [20]. The Min-Max normalization formula is as follows:

$$X\_scaled = (X - Xmin) / (Xmax - Xmin)] \quad (3)$$

The dataset's minimal value is denoted by Xmin, its maximum value by Xmax, and the original value is represented by X.

2.2.2 Normalization of Z-Score:

The values of a feature are scaled using this method to have a mean of 0 and a standard deviation of 1. To calculate this, take each number, remove the feature mean from it, and divide the result by the standard deviation. The following is the Z-Score normalization formula:

$$X\_scaled = (X - mean) / Standard\ Deviation \quad (4)$$

The original value is represented by X, the dataset mean is denoted by Mean, and the dataset standard deviation is represented by Standard deviation.

2.3 Geographical Information System (GIS)

Geospatial Information Systems, also known as Geographic Information Systems (GIS), became known in the early 1980s. A GIS is a system for obtaining, storing, analyzing, and managing spatial data along with data-related attributes that are remotely referenced to the Earth. Based on technology and its implementation, geographic information systems can be categorized into three applications:

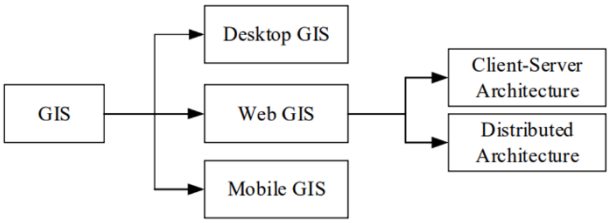

Fig. 2. Category Geographic Information Systems

3 Result and discussion

Figure 3 below provides a summary of the investigation.

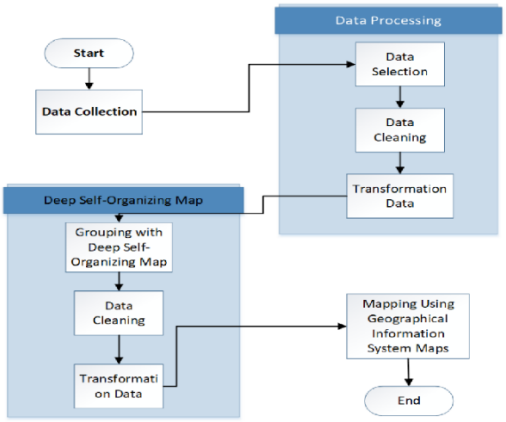

Fig. 3. Overview of the research

3.1 Collection of data

At this stage, it is done to collect the data that will be used in this research. This data is a type of secondary data, namely high school alumni data. The alumni data attributes to be used in this study include Student Number, student name, gender, address, no phone, study program, generation, year of graduation, final grade, place of work, address of workplace, and suitability of field of work.

3.2 Data processing

In this phase, there are three steps: data selection, data cleaning, and data transformation. The goal of this phase is to convert raw data into quality data, such as correcting incomplete (incomplete), noisy (data containing value errors), and inconsistent data.

- Data selection: In the data selection stage, we will select the attributes to be used. Of course, not all attributes are included in the dataset used in the data mining process because only attributes that become identification references will be selected. Of all the attributes mentioned above, the ones that will be used in the data selection process are Student Identification Number, Student Name, Gender, Major, and Final Grade.
- Data Cleaning: Before the data mining process can be implemented, the data cleaning process must be carried out. The purpose of this process is to ensure the quality of the data that has been selected at the data selection stage. Suppose gender writing is still not inconsistent there are still men and women using it.
- Data transformation: At this stage, the data will be transformed so that it is suitable for use in the clustering process. An attribute of the field of work is the categorical

data type Final Grade of the major. This process is carried out using the Self-Organizing Map method.

3.3 Self-Organizing Map

Z-Score and Min-Max are two of the three normalization procedures available in the Self-Organizing Map approach:

Clustering with a Self-Organizing Map using Z-Score and Min-Max normalization methods, aims to know data normalizations before forming SOM map training data. The results of normalization use Z-Score and Min-Max with the following datasets and header components in Table 1. As follows Information header Column:

1. NIS : Student Identification Number
2. TL : Place of birth
3. JK : Gender
4. KK : Skill Competency
5. AGM : Religion
6. PKN : Pancasila and Citizenship Education
7. BIN : Indonesian Language
8. MTK : Mathematics
9. SJI : History of Indonesia
10. BIG : English Language
11. SB : Arts and Culture
12. PJK : Physical Education, Sports and Health
13. BDH : Regional Language
14. UKK : Competency Test

Table. 1. Examples of Student Data

NIS	TL	JK	KK	Status	School Test Scores													UKK
					AGM	PKN	BIN	MTK	SJI	BIG	SB	PJK	BDH	SIMD	EKB	ADM	DP	
7130	1	1	1	1	93.5	94.5	93	87.5	91.5	93.5	93.5	90	91.5	93	91	89.5	92	98
7131	1	1	1	1	92.5	96.5	92	90.5	92	93.5	93.5	89	91	95	93	90	94	98
7132	1	1	1	1	90.5	92.5	85.5	84.5	90	88	88	89	90.5	88.0	85.0	87.5	90.5	90
7133	2	1	1	2	88	96.5	93.5	92.5	93.5	95.5	95.5	90	92.5	95.0	94.0	92.5	94.5	98
7134	3	1	1	1	92.5	95	88.5	84.5	88.5	92.5	92.5	88	91	90	91	91	90	91
7135	4	1	1	1	91.5	93.5	88.5	83.5	88.5	92	92	90	91	89	86	90	90	90
7136	2	1	1	1	92.5	92.5	87.5	83.5	87.5	94.5	94.5	85	91	92	88	88	91	92

NIS	TL	JK	KK	Status	School Test Scores													UKK
					AGM	PKN	BIN	MTK	SJI	BIG	SB	PJK	BDH	SIMD	EKB	ADM	DP	
7137	1	1	1	1	92.5	94	88.5	82.5	88.5	95.5	95.5	94	90	89	85	89	88	92

From an example of the initial dataset in Table 1. Normalization data such as Z-Score and Min-Max are produced, from which data is used to identify clustering data whose map size is specified as 10 x 10 and each cell holds 20 random weight values assigned. Table 2. Shows the initial weight sample for 1 cell:

Table 2. The Early Cell (1000 Train)

Data Weight																			
Min Max Initial Cell Weight	0.2	0.2	0.2	0.7	0.2	0.2	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1
Z-Score Initial Cell Weight	0.9	0.9	0.9	1.0	0.9	0.9	0.9	0.9	0.6	0.6	0.6	0.7	0.7	1.1	1.1	1.6	1.0	1.0	1.0

On the SOM algorithm for each iteration, the value of the node weight is adjusted using equation (1). In Table 3. List the final weight for 1 cell:

Table 3. End of cell (1000 Train)

Data Weight																			
Min Max Final Cell Weight	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.4	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.2
Z-Score Final Cell Weight	0.6	0.7	0.7	1.0	0.9	0.9	0.9	0.9	0.9	1.2	1.0	1.0	0.8	0.8	0.8	1.1	0.6	0.6	0.5

On the SOM algorithm for the initial weight of 1 cell on the train 2000, the value of the weight of the node can be seen in Table 4. Listing the starting weight for 1 cell:

Table 4. The Early Cell (2000 Train)

Data Weight																			
Min Max Initial Cell Weight	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.3	0.3	0.1	0.1	0.1
Z-Score Initial Cell Weight	0.6	0.5	0.8	0.9	0.9	1.8	1.3	1.3	1.0	0.9	0.9	1.2	1.2	1.3	1.0	1.0	0.8	0.7	0.9

On the SOM algorithm for each iteration, the value of the node weight is adjusted using equation (2). In Table 5. List the final weight for 1 cell:

Table 5. The end of cell (2000 Train)

Data Weight																			
Min Max Initial Cell Weight	0.1	0.1	0.2	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.3	0.3	0.5	0.1	0.1	0.1
Z-Score Initial Cell Weight	0.5	0.4	0.6	1.5	1.2	1.2	1.2	1.2	1.2	1.9	1.1	1.1	1.0	1.0	1.0	1.1	1.1	1.1	1.0

After 1000 trains and 2000 trains were performed, the results were then visualized with the U-Matrix of each normalization. Visual representation of Self-Organizing Maps is achieved primarily via the use of a color scale in the integrated distance matrix (U-Matrix). Fig. 4 shows that nearby nodes are represented identically and share the same hue, suggesting that they are part of the same cluster. This is a visualization of the Self-Organizing Matrix-U Map produced from 10x10 with 1000 trains as seen in the Fig. 5. Of the matrix -U produced by 10×10 with 2000 trains.

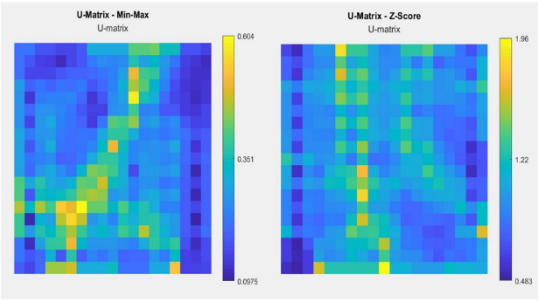

Fig. 4. U-Matrix 1000 Train

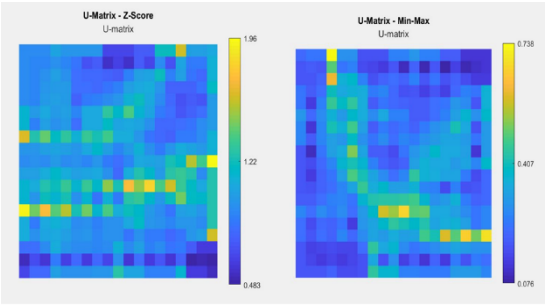

Fig. 5. U-Matrix 2000 Train

Analysis of the results of SOM clustering using Z-Score and Min-Max normalization by giving different train weights has an influence, namely the pattern of clustering when the train value is changed from 1000 trains to 2000 trains, when 1000 trains with two normalizations running shows there are two different clusters in the best cluster position with 5.31% Z-Score and 4.91% Min-Max, while when 2000 trains show different numbers but in the same cluster, with a value of 5.31% Z-Score and 3.98% Min-Max. The visualization results in Fig. 4. And Fig. 5. Which can be seen that some data falls into one cluster: the blue-colored cluster. In general, the map indicates that there are two to three clusters made up of one dominating cluster, while there are other clusters near the blue hue, especially yellow and orange in certain areas. This result is consistent with the conventional clustering of vocational high school students that categorizes status into 3 groups: working, continuing, and entrepreneurship. The major cluster is generally the employment cluster because the main goal of vocational high schools is to generate graduates who are ready to work, but in this analysis, there is a cluster that shows that those who continue to higher education are the most dominant shown in blue, then there is an entrepreneurial cluster that occupies the second dominant position shown in orange, and finally is employment shown in yellow.

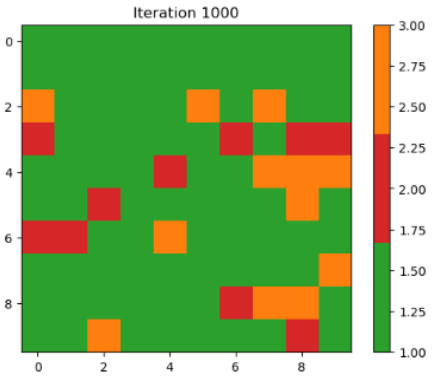

Fig. 6. SOM Label on the trained max iteration

To construct a label map, assign one label to each neuron mapped using random data. Figure 3 shows the label maps created for the first and last iterations. Initially, many neurons have values that are not 0 or 1 and the class labels appear randomly distributed, whereas, in the final iteration, there is a clear separation between classes 0 and 1 even though there are some cells that do not belong to one class in the last iteration, from the separation between class 0 and 1 produces an accuracy of 69%. The low accuracy value in Neural Networks with Self-Organizing Map (SOM) Architecture which uses the Min-Max algorithm is caused by many factors, therefore in the future, it is necessary to optimize parameters, add features, and develop more complex modeling so that system reliability and performance can be achieved improved.

4 Conclusion

This research has generated initial insights into grade data, and graduation status of student participants by identifying clusters of students within specific areas of expertise competencies. The total number of possible clusters is three, which correspond to the conventional categories of Working, Continuing, and Entrepreneurship. The cluster with the largest number of members is the Continuing cluster, and in this research, a self-organizing feature using a neural network is added to the Self-Organizing Map, thus providing a comparison where the normalization results produce an accuracy of 69% for the results from the last iteration. Future research plans to improve accuracy will be developed using Deep Self-Organizing Map.

References

1. T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, “SOM PAK: The Self-Organizing Map Program Package SOM PAK: The Self-Organizing Map Program Package,” 1996.
2. T. Kohonen, Biol. Cybern. **43**, 1 (1982)
3. S. H. Handoko, E. Sediono, S. Suhartono, J. Sist. Inf. Bisnis **1**, 2 (2014)
4. E. Sehirli, K. Arslan, Expert Syst. Appl. **205** (2022)
5. O. J. Alabi, J. W. Ng’ambi, D. Norris, Asian J. Anim. Vet. Adv. **7**, 2 (2012)

6. S. Lakho, A. Hussain Jalbani, M. S. Vighio, I. A. Memon, S. S. Soomro, Q.-N. Soomro, Sukkur IBA J. Comput. Math. Sci. **1**, 2 (2017)
7. F. T. Anggraeny, I. Y. Purbasari, E. Suryaningsih, *ReliefF Feature Selection and Bayesian Network Model for Hepatitis Diagnosis*, in 3rd Int. Conf. Inf. Technol. Bus. (2017)
8. A. Usman, O. L. Adenubi, J. Sci. Inf. Technol. **1**, 2 (2013)
9. K. L. Wu, M. S. Yang, Pattern Recognit. Lett. **26**, 9 (2005)
10. F. Höppner, F. Klawonn, Mathware & soft computing **11**, 5 (2004)
11. E. Yohannes, S. Ahmed, Int. J. Comput. Appl. **180**, 40 (2018)
12. M. W. Bara, N. B. Ahmad, M. M. Modu, H. A. Ali, *Self-organizing map clustering method for the analysis of e-learning activities*, In Majan international conference (MIC), IEEE (2018)
13. Y. Lee, J. Educ. Comput. Res. **57**, 2 (2019)
14. P. N. E. Nohuddin, Z. Zainol, A. Nordin, Zulfaqar Journal of Defence Science, Engineering & Technology **1**, 1 (2018)
15. D. Ifenthaler, I. Masduki, N. M. Seel, Instructional Science **39** (2011)
16. B. H. Kunaryo, M. Somantri and A. Sofwan, *Applicant Mapping of Predicting Seriousness in Choosing Private University using Self Organizing Map (SOM) Algorithm*, in 2020 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), Semarang, Indonesia (2020)
17. L. Chien-Sing, Y. P. Singh, *Student modeling using principal component analysis of SOM clusters*, in IEEE International Conference on Advanced Learning Technologies, Joensuu, Finland (2004)
18. S. R. M. de Campos, R. Henriques, *The knowledge discovery through the student's higher education census data*, in 13th Iberian Conference on Information Systems and Technologies (CISTI), Caceres, Spain (2018)
19. S. Delgado, F. Morán, J. C. S. José and D. Burgos, IEEE Access **9** (2021)
20. S. S. Ramakrishna, Int. J. Adv. Res. Comput. Sci. **9**, 2 (2018)