isemantic2021

by Yuni Yamasari

Submission date: 24-Sep-2021 09:46PM (UTC+0700)
Submission ID: 1656500228
File name: Students_Performance_using_Feature_Selection_isemantic_2020.pdf (3.56M)
Word count: 3488
Character count: 17987

Improving the Quality of the Clustering Process on Students' Performance using Feature Selection

Yuni Yamasari
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
yuniyamasari@unesa.ac.id

Anita Qoiriah
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
anitaqoiriah@unesa.ac.id

Hapsari P. A. Tjahyaningtijas
Department of Electrical Engineering
Universitas Negeri Surabaya
Surabaya, Indonesia
hapsaripeni@unesa.ac.id

Ricky E. Putra
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
rickyeka@unesa.ac.id

Agus Prihanto
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
Agusprihanto@unesa.ac.id

Asmunin
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
asmunin@unesa.ac.id

***Abstract*— the quality of students' performance clusters relates to the accuracy of students being in groups based on their performance. However, the resulting quality sometimes needs to be improved because the clustering process involves features that are not dominant. Furthermore, in the previous works, measurement of the quality of the clusters in unsupervised evaluation often only uses one measure. Therefore, this paper focuses to enhance the quality of clusters by eliminating features that are irrelevant by applying the feature selection method called the Gini Index. Meanwhile, in this paper, the clustering method applied is K-means for the mining process. Then, we propose the evaluation process measured by three metrics, namely: silhouette coefficient, ANOVA, and t-test. The experimental results show that the Gini Index can improve the quality of clusters based on the three proposed metrics. Keywords—clustering, students' performance, feature selection, t-test, Gini Index, K-means**

I. INTRODUCTION

Nowadays, almost all processes in the world of education utilize the advancement of information and communication technology [1] with its security aspect [2]. This condition has the consequence obtained large amounts of data. One of them is student data. This data encourages researchers to research in the field of Educational Data Mining (EDM). One method that is quite often done is clustering [3].

Relating to clustering, researchers apply it to many domains. Najdi et al do the clustering on the student typologies[4]. Then, Cerezo et al exploit it to analyze the students' LMS interaction patterns[5]. Harwati et al use the clustering method on the mapping of the students' performance [6]. Singh et al also do the clustering on students 'performance [7].

However, the clustering process still involved all the features of these previous studies. Whereas, these features often contain irrelevant features which can cause a decrease in the quality of the resulting clusters. For this reason, some researchers have sought to improve the quality of clusters by choosing only dominant features. Feature selection is applied on the psychomotor domain [8], the hybrid feature selection is exploited on the students' achievement [9], features selection also is used to improve the sentiment analysis on the teaching evaluation [10], and features of data set relating to e-course are selected in [11] to evaluate the e-learning process, etc.

Gini is one of the feature selection methods that are not much explored in EDM[12]. Gini 'is a modification of another attribute's quality feature called the Gini index. Both the Gini Index and Gini' are the same as information gain in terms of bias towards the importance of many high-value attributes. So the Gini Index as one of the feature selection methods needs to be explored in the EDM domain.

Regarding the measurement of the quality of the cluster, previous studies use the supervised evaluation [9]. Other studies use the unsupervised evaluation [8], [13], [7]. In connection with the unsupervised evaluation, previous studies generally use a single measure.

Therefore, this research focuses on the clustering process which only involves relevant or dominant features by applying the feature selection method. This application is intended so that the cluster produced has high quality. Furthermore, we propose cluster quality measurements in several metrics, namely: silhouette coefficient, ANOVA, and t-test.

Finally, our paper is specified for 4 sections: (1) introduction, (2) the proposed method (3) result and, (4) conclusion.

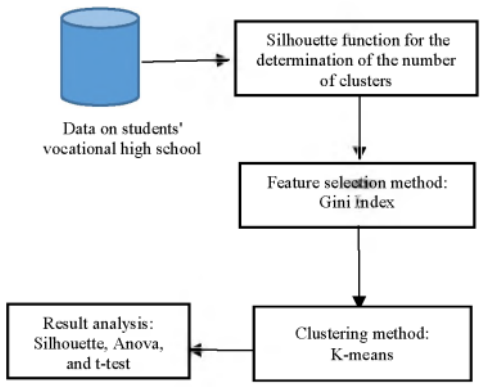


Fig. 1. The proposed method

978-1-7281-9068-6/20/$31.00 ©2020 IEEE

II. METHOD

This section describes the steps of the proposed method. Five steps included the student data as shown in Fig.1.

Step 1: Student data collected from the Vocational High School Surabaya

The exploration of this research uses student data on previous research [14]. The student data is stored when the students interact with the e-learning system, especially online testing. This process generates 101 features and then features are extracted to produce the simple features. This activity obtains 5 features. They are as follows: Done, PercentTrue, Time, Hint, and Score which having the numeric data type.

Step 2: determining the optimum number of clusters using the Silhouette function

The function used in this step is called the silhouette coefficient. This function is also exploited in the evaluation step as the last step. The formula of this function is represented in equation (2).

Step 3: doing the feature selection method

In this step, the Gini index as a feature selection method is applied to our student data. The goal of this step is to find the appropriate features for the clustering process.

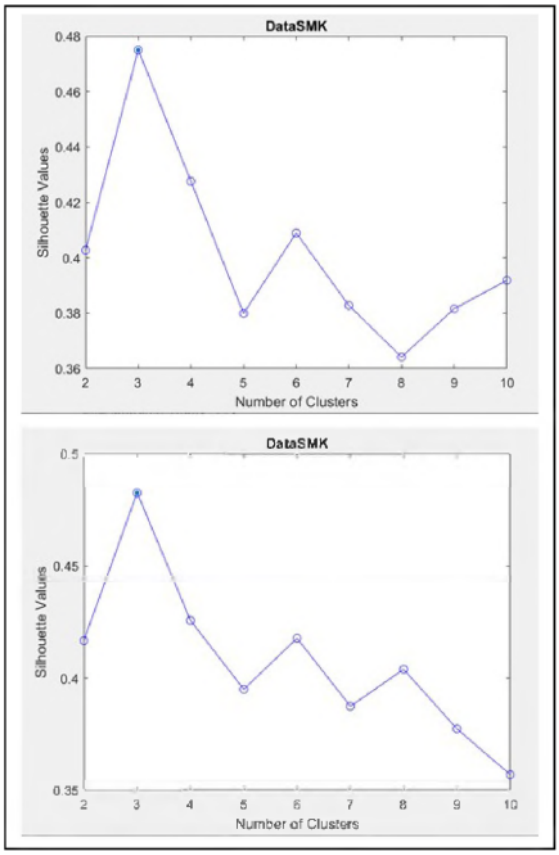
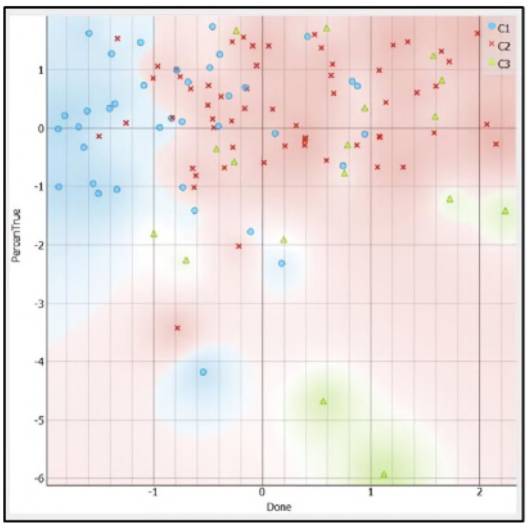
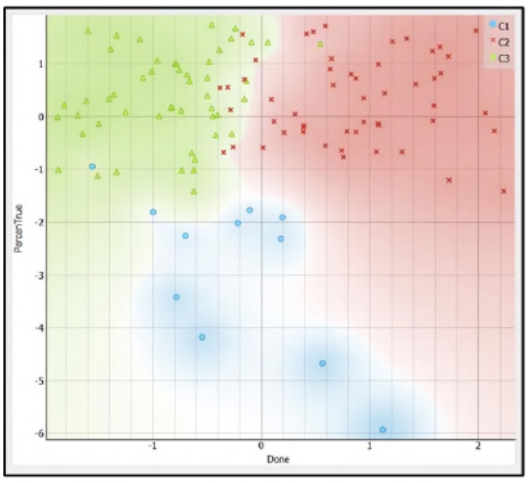


Fig. 2. Silhouette value which shows the optimal number of clusters in 2 10 times running



(a).



(b)

Fig. 3. The visualization clustering results in (a). original feature sets & K-means on, (b). selected feature sets & K-means

So the process does not involve the irrelevant features which can decline the clustering performance. The formula of this feature selection method is as follows:

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p_i]^2 \quad (1)$$

Where the Gini index measures the purity of feature t toward category c. p_i is the probability of an instance that is classified to a particular class i [15][16].

Step 4: Applying the clustering method

For the clustering process, our research explores the popular clustering method, namely: K-means [3]. Almost

previous works in Educational Data Mining (EDM) apply it to do the clustering process

Step 5: Analyzing the results using silhouette coefficient, Anova, and t-test

To evaluate clusters, this research applies the unsupervised evaluation which only uses the internal information. In this step, we explore the silhouette value [17] on every point [18]. This is the ratio of points matched with other points in a cluster when faced with points in another cluster. In this context, the degree of similarity depends on the distance between points. The more similar a point is to another point, the closer one point is to another point. The silhouette value of ith point, S_i, is expressed in equation (2).

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad (2)$$

Where, a_i is the average distance calculated between the i-th point and other points on cluster i. b_i Also, it represents the minimum distance measured between points to other points in another cluster. The silhouette value is from -1 to +1, where a higher value means that the i-th point is more suitable for the cluster than for other clusters.

TABLE I. ANALYSIS OF CLUSTER RESULTS IN EACH COMBINATION WITH ANOVA

Silhouette value on the combination of original feature sets and K-means			Silhouette value on the combination of selected feature sets and K-means		
C1	C2	C3	C1	C2	C3
0.584	0.568	0.546	0.503	0.525	0.676
0.511	0.59	0.54	0.625	0.541	0.673
0.541	0.572	0.587	0.647	0.556	0.634
0.584	0.511	0.561	0.639	0.555	0.677
0.52	0.566	0.522	0.548	0.582	0.666
0.621	0.603	0.544	0.468	0.554	0.678
0.612	0.583	0.503	0.486	0.603	0.679
0.603	0.573	0.602	0.535	0.613	0.684
0.619	0.606	0.561	0.579	0.622	0.681
0.616	0.573	0.568	0.637	0.588	0.681
0.55	0.616	0.594	0.609	0.637	0.637
0.485	0.626	0.562		0.649	0.681
0.481	0.606	0.583		0.655	0.687
0.551	0.568	0.525		0.66	0.667
0.596	0.574	0.501		0.664	0.681
0.601	0.569	0.59		0.657	0.684
0.492	0.565	0.487		0.672	0.681
0.483	0.642			0.675	0.664
0.609	0.645			0.678	0.685
...	...		...	...	

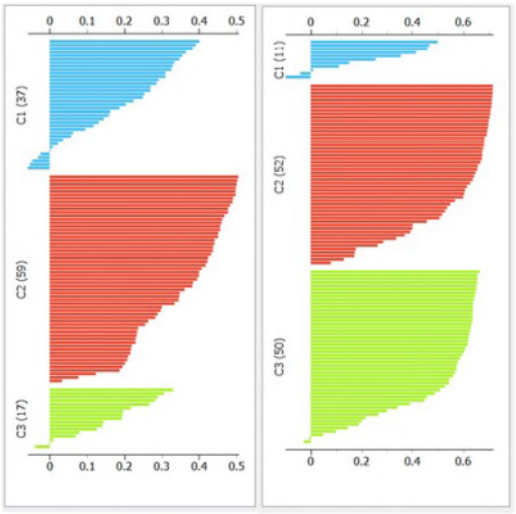


Fig. 4. The visualization of the silhouette value on (a) the original K-means feature set and on (b) the selected feature set K-means

More than that, we also conducted a clustering analysis with ANOVA as another evaluation. In this study, ANOVA was used to investigate whether the clusters produced by each method differed significantly based on the value of the silhouette. Finally, the evaluation was carried out using a t-test to see the quality of the cluster between 2 methods based on the value of the silhouette.

III. RESULT

In this section, the execution of the proposed framework is investigated. We describe this section in 3 sub-sections, namely: the optimum number of clusters, clustering process using K-means, and evaluation of the result.

A. The optimum number of cluster

After the raw features are extracted, the silhouette function is run as the number 10 times on student data obtained. The result value is represented in the array of the silhouette value, as follows: [NaN 0.4167 **0.4826** 0.4257 0.3950 0.4178 0.3874 0.4039 0.3774 0.3569]. This array shows that the optimum value is 3 clusters. Further, this array is plotted in a graph which indicates that the optimum number of clusters on student data is three clusters. So we determine the number of clusters = 3 for the next step as shown in Fig.2.

Next, this research selects features using the Gini Index. The applied of this method generates two features, namely: done and precentTrue. This step has a goal to increase the quality of clusters and its result is mined with the popular clustering method called "K-means".

B. The clustering process using K-means

Here, the clustering process uses K-means applied on two features sets, namely: original feature set and the selected feature set. The silhouette value is described in Table II. The

visualization of the clustering result is depicted in Fig.3. It is found that the K-means on the selected feature set has the cluster quality better than on the original feature set. This quality is based on the intra- and between-cluster. The intra-cluster is related to closeness in one cluster and between-cluster is related to the distance between clusters.

C. The evaluation of the result

In the final step, we analyze the results with 3 measurements, namely: silhouette, ANOVA, and t-test. The silhouette value itself can be used to see the validity of the resulting cluster, as explained earlier. The fewer silhouette values below 0, the higher the validity of the cluster. Because the value of 0 indicates the cognitive level of students is between 2 clusters and the value - (negative) on the silhouette value which indicates the unsuitable of students being in the cluster. The silhouette results show that the validity of the K-means combination and the selected feature set is higher than the K-means combination on the original feature set. This is because the number of silhouette values is smaller than 0 on the combination of selected feature and K-means is less than on the combination of the original feature set and K-means as shown in Fig.4.

Furthermore, we evaluate the results of the clusters by using ANOVA to investigate whether the clusters produced by each method differ significantly based on the value of the silhouette. So, the evaluation using ANOVA is done twice, the first: cluster results on a combination of selected feature sets and k-means, and the second: cluster results on a combination of original feature sets and k-means.

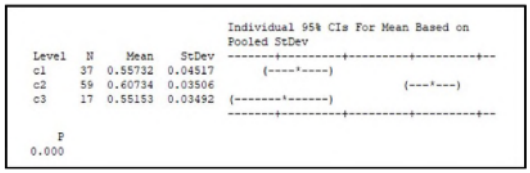


Fig. 5. The cluster analysis results from a combination of K-means and original feature set with ANOVA

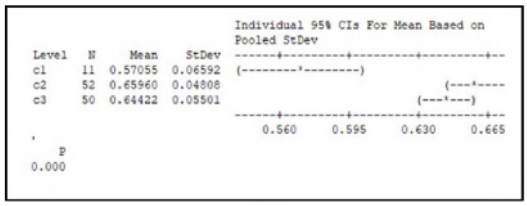


Fig. 6. The Cluster analysis result from a combination of K-means and feature sets selected with ANOVA

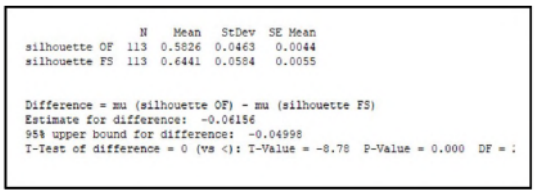


Fig. 7. T-test of the means silhouette values on both methods

TABLE II. THE SILHOUETTE VALUE OF THE COMBINATION OF THE ORIGINAL K-MEANS & FEATURE SET (OF) AND THE COMBINATION OF THE SELECTED K-MEANS & FEATURE SET (FS)

silhouette OF	*silhouette* FS
0.584	0.503
0.511	0.625
0.541	0.647
0.584	0.639
0.52	0.548
0.621	0.468
0.612	0.486
0.603	0.535
0.619	0.579
0.616	0.637
0.55	0.609
0.485	0.525
0.481	0.541
0.551	0.556
0.596	0.555
0.601	0.582
0.492	0.554
0.483	0.603
0.609	0.613
0.537	0.622
0.578	0.588
0.507	0.637

The hypothesis of the ANOVA test is as follows: the average based on the silhouette value of the clustering results on the combination of K-means and the original feature set and the selected feature set are:

H0: there is no difference in the means of silhouette values in all clusters produced (means/average of Silhouette C1 = means of Silhouette C2 = means of Silhouette C3)
H1: there is a difference from the means of silhouette values in all the clusters produced
(means of Silhouette C1 ≠ means of Silhouette C2 ≠ means of Silhouette C3)
With the determination of the confidence level = 95% and α = 0.05. This means that if P-Value < α then reject H0.
Cluster analysis results from a combination of K-means and original feature sets are presented in Fig.5. Members of cluster c1, c2, c3 are 37, 59, and 17 students, respectively. here, the resulting P-Value is 0, so reject H0 and accept H1. This means that there is a significant difference from the silhouette values in the cluster generated by the combination of K-means and the original features set as shown in Fig. 5.

Meanwhile, the clustering analysis results from the combination of K-means, and selected feature sets are presented in Fig.6. The members of each cluster c1, c2, c3, respectively, are 11, 52, and 50 students. The resulting P-Value is 0, so reject H0 and accept H1. This means that

there is a significant difference from the silhouette values in the cluster produced by the combination of K-means and the selected feature set. So the cluster quality of the two combinations is good because the clusters have a significant difference in terms of the means of silhouette values as shown on Fig.6..

Finally, we analyze the quality of clustering results from both methods. We performed a t-test on the average of silhouette values of the two methods (a combination of K-means & original feature set (OF) and a combination of selected K-means & feature sets (FS)) as shown in Table II. The hypothesis of this t-test is as follows: based on the average of silhouette values of the two methods is as follows depicted in Fig.7:

H0: means of silhouette value_OF ((the combination of K-means & original feature set)> means of silhouette value_FS (the combination of K-means & selected feature set).

This means that (means of Silhouette OF> means of Silhouette FS)

H1: means of silhouette value OF ((the combination of K-means & original feature set) < means of silhouette value_FS (a combination of K-means & feature set selected

This means that (average Silhouette OF <average Silhouette FS).

With a confidence level of 95% and $\alpha = 0.05$ and the alternative is "less than". This means that if the P-Value < α then reject H0. T-test results show that the P-value = 0. This means that reject H0 and accept H1 as presented in Fig.7. Thus, the combination of the K-means & the selected feature set (FS) is of higher quality than the combination of the K-means & original feature set (OF) because of (average Silhouette OF < average Silhouette FS).

IV. Conclusion

The selection of features using the Gini Index can increase cluster quality. The measurement results show that students who have low silhouette levels are reduced. This is also supported by the results of investigations from the t-test and ANOVA which illustrate the same results that feature selection can increase silhouette levels in students.

Acknowledgment

The authors would like to thanks the Department of Informatics, Universitas Negeri Surabaya, Indonesia for support.

References

[1] L. Juhaňák, J. Zounek, and L. Rohlíková, "Using process mining to analyze students' quiz-taking behavior patterns in a learning management system," *Comput. Human Behav.*, Dec. 2017.

[2] P. Maniriho and T. Ahmad, "Information hiding scheme for digital images using difference expansion and modulus function," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 31, no. 3, pp. 335–347, Jul. 2019.

[3] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, Mar. 2014.

[4] L. Najdi and B. Er-Raha, "Implementing cluster analysis tool for the identification of students typologies," in *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, 2016, pp. 575–580.

[5] R. Cerezo, M. Sánchez-Santillán, M. P. Paule-Ruiz, and J. C. Núñez, "Students' LMS interaction patterns and their relationship with achievement: A case study in higher education," *Comput. Educ.*, vol. 96, pp. 42–54, May 2016.

[6] H. Harwati, A. P. Alfiani, and F. ayu Wulandari, "Mapping Student â€™s Performance Based on Data Mining Approach (A Case Study)," in *Procedia - Social and Behavioral Sciences*, 2015, vol. 38, no. 1, pp. 173–177.

[7] I. Singh, A. S. Sabitha, and A. Bansal, "Student performance analysis using clustering algorithm," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, 2016, pp. 294–299.

[8] Y. Yamasari, S. M. S. Nugroho, R. Harimurti, and M. H. Purnomo, "Improving the cluster validity on student's psychomotor domain using feature selection," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 460–465.

[9] Y. Yamasari, S. M. S. Nugroho, K. Yoshimoto, H. Takahashi, and M. H. Purnomo, "Identifying Dominant Characteristics of Students' Cognitive Domain on Clustering-based Classification," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 1, 2020.

[10] C. Pong-Inwong and K. Kaewmak, "Improved sentiment analysis for teaching evaluation using feature selection and voting ensemble learning integration," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, 2016, pp. 1222–1225.

[11] M. A. Hogo, "Evaluation of e-learning systems based on fuzzy clustering models and statistical tools," *Expert Syst. Appl.*, vol. 37, no. 10, pp. 6891–6903, 2010.

[12] M. Zaffar, M. A. Hashmani, and K. S. Savita, "Performance analysis of feature selection algorithm for educational data mining," in *2017 IEEE Conference on Big Data and Analytics (ICBDA)*, 2017, pp. 7–12.

[13] L. S. Robles Pedrozo and M. Rodriguez-Artacho, "A cluster-based analisys to diagnose students' learning achievements," in *2013 IEEE Global Engineering Education Conference (EDUCON)*, 2013, pp. 1118–1123.

[14] Y. Yamasari, S. Mardi, S. Nugroho, K. Yoshimoto, H. Takahashi, and M. H. Purnomo, "Expanding Tree-Based Classifiers Using Meta-Algorithm Approach: An Application for Identifying Students' Cognitive Level," *Int. J. Innov. Comput.*, vol. 15, no. 6, pp. 2085–2107, 2019.

[15] J. Yang, Z. Qu, and Z. Liu, "Improved Feature-Selection Method Considering the Imbalance Problem in Text Categorization," *Sci. World J.*, vol. 2014, 2014.

[16] G. Shobha and S. Rangaswamy, "Gini Index - an overview | ScienceDirect Topics," *scincedirect*, 2009. [Online]. Available: https://www.sciencedirect.com/topics/mathematics/gini-index. [Accessed: 02-Jul-2020].

[17] P. J. Rousseeuw, "Silhouettes : a graphical aid to the interpretation and validation of cluster analysis," vol. 20, pp. 53–65, 1987.

[18] A. Starczewski and A. Krzyżak, "Performance Evaluation of the Silhouette Index," Springer, Cham, 2015, pp. 49–58.

isemantic2021

ORIGINALITY REPORT

8%	6%	4%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Oklahoma Health Science Center Student Paper	2%
2	arsip.its.ac.id Internet Source	1%
3	Armin Moghimi, Safa Khazai, Ali Mohammadzadeh. "An improved fast level set method initialized with a combination of k-means clustering and Otsu thresholding for unsupervised change detection from SAR images", Arabian Journal of Geosciences, 2017 Publication	1%
4	0-www-crossref-org.libus.csd.mu.edu Internet Source	<1%
5	mafiadoc.com Internet Source	<1%
6	"Recent Trends in Data Science and Soft Computing", Springer Science and Business Media LLC, 2019 Publication	<1%

7	Abdallah Moubayed, MohammadNoor Injadat, Ali Bou Nassif, Hanan Lutfiyya, Abdallah Shami. "e-Learning: Challenges and Research Opportunities Using Machine Learning & Data Analytics", IEEE Access, 2018 Publication	<1 %
8	iopscience.iop.org Internet Source	<1 %
9	ir.lib.uwo.ca Internet Source	<1 %
10	revistasojs.utn.edu.ec Internet Source	<1 %
11	"Information and Communications Security", Springer Science and Business Media LLC, 2020 Publication	<1 %
12	hoques.com Internet Source	<1 %
13	Mari-Sanna Paukkeri, Jaakko Väyrynen, Antti Arppe. "Chapter 1 Exploring Extensive Linguistic Feature Sets in Near-Synonym Lexical Choice", Springer Science and Business Media LLC, 2012 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		