CLUSTERING SUBJECTS IN LAMPUNG PROVINCIAL NATIONAL EXAMINATION OF JUNIOR HIGH SCHOOL THROUGH MAXIMUM SPANNING TREE

Sugama Maskar[1], Nicky Dwi Puspaningtyas[2*]

[1,2*]Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
*Jl. ZA. Pagar Alam No.9-11, 35132, Bandar Lampung, Indonesia
E-mail: sugama_maskar@teknokrat.ac.id [1)]
            nicky@teknokrat.ac.id [2*)]

Abstrak

Ujian Nasional yang dilaksanakan setiap tahun di Indonesia telah menghasilkan banyak data, termasuk di Provinsi Lampung. Pada rentang tahun 2015 sampai dengan tahun 2018, telah terkumpul kurang lebih 11 juta data hasil ujian nasional. Data tersebut tentu dapat memberikan banyak informasi untuk perbaikan Pendidikan Indonesia di masa mendatang. Tulisan ini membahas tentang pengelompokan mata pelajaran pada Ujian Nasional SMP di Provinsi Lampung dengan tujuan untuk memetakan hubungan hasil belajar Matematika, IPA, Bahasa Indonesia, dan Bahasa Inggris. Tujuan pengelompokan tersebut sebagai analisa lebih jauh agar perbaikan pendidikan juga dapat mulai masuk pada ranah non-teknis, seperti penjadwalan, penyeusaian waktu belajar di kelas sampai dengan waktu pengerjaan tugas, serta perbaikan pada penentuan mata pelajaran yang menjadi syarat pelajaran lainnya. Analisis klusterisasi dilakukan dengan menggunakan Algoritma Kruskal berdasarkan *graf maximum spanning tree* (MST) dengan bantuan teknik statistik deskriptif dan inferensial. Data yang digunakan adalah hasil Ujian Nasional di Provinsi Lampung tahun 2017 dan 2018 sebanyak 15876 data. Graf MST ditentukan menggunakan algoritma Kruskal dan koefisien korelasi ditentukan menggunakan nilai koefisien korelasi Pearson. Hasil eksperimen menunjukkan bahwa Matematika, IPA, dan Bahasa Inggris memiliki korelasi yang paling kuat. Mata pelajaran yang paling kuat korelasinya dengan Bahasa Indonesia adalah Bahasa Inggris. Oleh karena itu, pemetaan mata pelajaran dapat disusun menjadi dua kelompok yaitu Matematika-IPA-Bahasa Inggris dan Bahasa Inggris- Bahasa Indonesia- IPA. Hasil tersebut dapat bermanfaat sebagai dasar pengambilan keputusan pada berbagai aspek teknis untuk mengoptimalkan hasil belajar siswa SMP khususnya di Provinsi Lampung.

Kata kunci: Klusterisasi, Lampung, Ujian Nasional

Abstract

National exams carried out every year in Indonesia have produced a lot of data, including in Lampung Province. In the span of 2015 to 2018, approximately 11 million national exam results have been collected. This data can certainly provide a lot of information for the improvement of Indonesian education. This paper discusses the clustering of subjects on junior high school national exams in Lampung Province with the aim of mapping the relationship of learning outcomes in Mathematics, Sciences, Indonesian, and English. The purpose of the clustering is as a further analysis so that educational improvements can also begin to enter the non-technical realm, such as scheduling, adjusting class and assignment time, and also determine prerequisite subjects. Clustering analysis is performed using an algorithm based on maximum spanning tree (MST) graphs with the help of descriptive and inferential statistical techniques. The data used are the results of national exams in Lampung province in 2017 and 2018 totaling 15.876 data. The MST Graph is determined using Kruskal's algorithm and the coefficient of correlation is determined using Pearson coefficient of correlation value. The experimental results show that Mathematics, Natural Sciences, and English have the strongest correlation. Subject that has the strongest correlation with Indonesian is English. Therefore, the mapping of subjects can be arranged into two groups namely Mathematics- Sciences- English and English-Indonesian-Sains. These results can be useful as a basis for decision making on various technical aspects to optimize junior high school student learning outcomes, especially in Lampung Province.

Keywords: Clusterization, Lampung, National Examination

INTRODUCTION

One of the evaluations of learning from the Indonesian government is national examination. Based on the official the Ministry of Education and Culture of Indonesia website, the education assessments center discharge annual analysis of the national examination's results. Those results are set as the basis for improving the quality of education. Based on data from www.unbk.kemendikbud.go.id the school who already took in the national test of computer based in 2015, 2016, and 2017 respectively 555, 4382, and 30.557 (Admin, 2017). That number does not include the schools that have taken national examination on paper. According to www.jpp.go.id that by 2018 the number of schools that took the national examination of computer based into 69.113 with a predicted number of participants 6.291.323 students. Based on this, over 11 million national examination data has been estimated to be collected from 2015 to 2018 as early as elementary school until senior high school. So far, the analysis of the national examination data has only been carried out with the main objective of mapping the students cognitive abilities in Indonesia which is measured based on descriptive statistics numerical extract from data processing on national examination scores (Maskar, 2020). Because there are so many data on national examination results, it would be a shame if the analysis is limited only to find out the average results of national examination scores. This data can actually provide much more in-depth and useful information for the improvement of education in Indonesia if analyzed in depth and precisely.

Processing and analysis techniques have developed significantly one of them is educational data mining (EDM). EDM is a subset of data mining which is a new area in processing educational data by designing and using appropriate algorithms. This area has great potential to produce a variety of benefits in several sectors in the field of education, including as decision making, guidance, giving criticism to students, teachers or scientists, improving administration of learning items, etc. that can be done by the authorities of educational institutions (Hussain et al., 2019; Mishra et al., 2017; Silva & Fonseca, 2017).

The area of EDM consists of data visualization, statistical analysis, data training, machine learning, grouping or clustering, classification, and outlier analysis (Kumar & Pal, 2011; Shingari et al., 2017). The main purpose of EDM is to make large volumes of dataset into useful information and can be used as a basis for decisions making by educational institutions. The method that might be used is a data clustering technique that has been successfully applied in data mining (Cheng et al., 2014). Data clustering can be done with several techniques or algorithms including single link, average link, minimum spanning tree, k-means(Li et al., 2019; Yang, 2017). Data clustering aims to extract relationships in a dataset and to determine interesting patterns based on sameness of the sample by grouping several data objects into groups or clusters so that objects in a cluster have a high degree of similarity and very different from objects in other clusters. Clustering techniques have been studied and applied in the fields of statistics, machine learning, pattern recognition and image processing (Cheng et al., 2014; Li et al., 2019; Yang, 2017). The contribution of this paper is to determine the relationship between junior high school national examination

outcomes in 2017 and 2018 for each subject in Lampung Province by clustering each subject based on the average national examination scores for each school. The Clustering is based on a modified graph method of the minimum spanning tree that is the maximum spanning tree (MST). The minimum path and spanning tree problems are two things that are similar and can be modified according to conditions. Minimum spanning tree method is the recommended graph method to clustering problem (Cheng et al., 2014; Song et al., 2019).

Referring to the results of the 2017 and 2018 junior secondary school national examination in Lampung Province through the www.hasilun.puspendik.kemdikbud.go.id page, there is a fact that the average score in 2017 is 50.88 with a comparison of the national average of 54.25, and in 2018 it is 40.87 with a national average of 51.10. This fact shows information that the average value of the junior high school national examination in Lampung Province is still below the national average. It is interesting that when referring to the national average data, the average value does not even reach 55. This shows that there is a problem with cognitive abilities, at a minimum, not only in Lampung Province, but also nationally.

These educational problems are not only found in the matter of the substance or non-technical aspects, but also on the technical aspects, including scheduling, determining the length of time to study in class, setting assignments for students, and other non-technical aspects related to the teaching and learning process. The curriculum change to *"Kurikulum 2013"* certainly has an effect on this aspect as well. These changes also indirectly change

the non-technical aspects of the learning process. This is often overlooked in the improvement of education problems Indonesia. The weight of improvement on technical or substance factors still dominates compared to non-technical factors. Therefore, analysis related to factors non-technical problems needs to be popularized and started to be carried out.

Previously Nurviana(2016) and Maskar (2020) have conducted similar research in different places using national examination score data on senior high school level. Nurviana work earlier in one school in Aceh with the results showing mathematics being the center of subjects which means that other subjects follow mathematical learning patterns. Furthermore, Maskar conducts analysis on national examination dataset for sains program in Lampung Province with the results of science and mathematics becoming the center of subjects. The analysis conducted in the two studies used a maximum spanning tree graph model.

Clasterization of subjects can be used as a basis for decision making by educational institutions for some technical matters. The aim is to optimize the learning outcomes of junior high school students, especially in Lampung Province. These technical issues include competency mapping in subjects, preparation of lesson schedules, distribution of subject learning hours and other policies. Effective and efficient policy formulation is one of the keys to the success of the institution to achieve the stated learning goals. Therefore, educational data mining is fundamental as the basis for making decisions to develop policies in educational institutions.

METHOD
Research Procedure

This study uses a quantitative research design by processing data on the results of the 2017 and 2018 junior high school national examination with data processing procedures starting with descriptive statistical methods through scatterplot analysis, followed by inferential statistical analysis through correlation coefficient values, and finally using the maximum spanning tree graph model. The three analyzes are then combined to draw conclusions regarding the grouping of subjects according to the abilities of junior high school students in that year.

Dataset

The data of the Lampung province junior secondary school in 2017 and 2018 were obtained from the Lampung provincial education office and website page www.hasilun.puspendik.kemendikbud.g o.id. The data consist of 1,994 and 1,975 schools in Lampung provinces in 2017 and 2018 respectively. Total data collected are 15,876 data consisting of four subjects; Science, Mathematics, Indonesian Language, and English.

Formation of the MST Graph Model

The subject clustering process is done using the maximum spanning tree (MST) graph method. This method is a modification of the minimum spanning tree. Modifications are made because weight of the graph for clustering subjects used correlation coefficient values, because if the correlation coefficient higher (maximum 1), then the correlation between the subjects was better. Minimum spanning tree is a sub-graph that does not have a cycle that connects all graph vertices with a minimum weight so that the graph is a graph model that is suitable for determining the relationship between vertices on a graph. The strongest relationship is vertex that has the most branches or degrees (Crobe et al., 2016; Kenneth H. Rosen, 2013).

Minimum spanning tree is one of the graph models that is made using Prim or Kruskal algorithm. The minimum spanning tree is built by all data by giving a weight to each edge of the graph and a number of steps to form a new graph by removing the edge that is greater than the previous weight.

The basic idea is to calculate sum of all the weights on the initial graph and begin with some vertices V on graph given $G = (V,E)$. All possible vertices that are directly connected to the starting point are taken in stages and weights of graph are reduced from the total one by one(Lingaya et al., 2019; Walter & Dubey, 2016).

Maximum Spanning Tree (MST) graph model in the subject clustering technique is formed by interpreting the graph point as a subject, there are four subjects namely Science, Mathematics, Indonesian Language and English. Furthermore, edge of the graph is interpreted as correlation between two subjects with graph weight is the value of correlation coefficient. The correlation coefficient value is obtained by using the Spearman correlation coefficient. As a result, an initial graph can be formed which is a complete graph as a model of the relationships between subjects in the national junior high school exams based on the value of the correlation coefficient as shown in Figure 1.
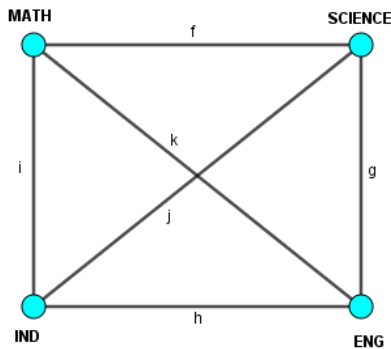
Figure 1. Initial Graph of Subject Relations in junior high school national examinations

Each vertex is interpreted as a subject where MATH means mathematics, SCIENCE is science, ENG is english, and IND is indonesian. The values of f,g,h,i,j, and k are the correlation coefficient values between subjects. Based on the initial graph, the MST graph is formed using the Kruskal's Algorithm from the largest to the smallest coefficient correlation based on the following provisions (Sugiyono, 2015):

Table 1. Level of Relationship Coefficient Correlation

Coefficient interval	Correlation level
0,00 – 0,199	Very weak
0,20 – 0,399	Weak
0,40 – 0,599	Moderate
0,60 – 0,799	Strong
0,80 – 1,00	Very strong

Based on Wattimena & Lawalatta (2013), with a slight modification, the formation of the Maximum Spanning Tree graph with the Kruskal's algorithm is as follows:
1) The weight edge of the graph G is sorted from the largest weight to the smallest.

2) Select the edge u ,v which has the maximum weight that does not form the cycle. Add u ,v into T.
3) Repeat step 2 until the Maximum Spanning Tree is formed, that is, when the side in the tree stretches T. number n-1 (n is the number of vertices in the graph).

The clustering process will be determined based on several analyzes including; scaterpolot graphs on descriptive statistical techniques, based on correlation coefficient values, and finally based on neighborhood relations on MST graphs. Analysis is carried out carefully so that the subject cluster can be formed in accordance with the most close relationships and in accordance with good clustering rules.

RESULTS ANDDISCUSSION

Results

Basic Analysis

This basic analysis aims to find out initial information from the dataset of the national examination results of Lampung Province in 2017 and 2018. The data is processed using SPSS software. Table 2 shows numerical esence of data processing as initial information for further analysis.

Table 2. Numerical Essence of Lampung Province Junior High School Dataset in 2017 and 2018

Item	2017	2018
Minimum	25.63	21.00
Maximum	93.36	86.03
Mean	49.40	45.01
Standards Deviation	12.67	12.57

Table 2 shows that there is a considerable range between the largest and smallest values of the two data

groups. The range of the first data group, in 2017, is 67.73 and in 2018 it is 65.03. The two data groups also have relatively the same mean and standard deviation, 49.40 compared to 45.01 for the mean and 12.67 and 12.57 for the standard deviation.

There are two important information that can be drawn from the average value and standard deviation of the two data groups. First, the average value of the two data sets is relatively low, below 50. The standard deviation of the two data shows that the average value illustrates the majority of each data item on the results of the national exam which means that the value supports information about the mean values previously mentioned. Second, the two data groups tend to be identical by looking at each item's value in the numerical score in table 2. Therefore, the process of analyzing the subject clustering tends to be easier because the two data groups do not have significant differences even though they are taken in different years.

Scatterplot Analysis

Scatter plot is a graph used to analyze the relationship or linearity between two groups of data. The purpose of the scatterplot analysis is to analyze correlations between subjects using a graphical approach. Following is the scatterplot of the national exams data in Lampung Province in 2017 and 2018.

Figure 2 shows that in 2017, subjects that have a close relationship are Science and English, Science and Mathematics, Mathematics and English. While Indonesian has almost no relationship with other subjects. In 2018 the relationship is shown by subjects in English and Natural Sciences, Mathematics and Natural Sciences, English and Mathematics. The

relationship between Indonesian and Science subjects, Mathematics and English is relatively the same, there tends to be a correlation but it is not as good as the relationship of other subjects. But if you pay close attention, Indonesian tends to have a good correlation with English.
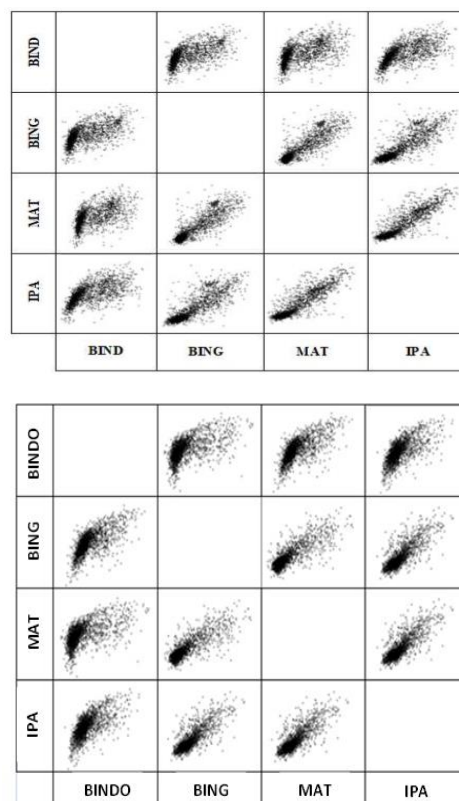

Figure 2. Scatterplot Relations between Subjects in the Lampung Province junior high school national examinations in 2017 and 2018

The analysis above shows that the strong relationship with the results of the national exams in Lampung Province in 2017 and 2018 is shown by three subjects, namely Natural Sciences, Mathematics and English. While Indonesian has a correlation with English. This shows the initial description of the subject mapping based on the strongest relationship where the subject mapping is divided into two groups and English is

contained in that two groups, while Indonesian is in one group with English.

Figure 3 shows the initial mapping description based on the results of the scatterplot analysis. The initial cluster in Figure 3 forms the basis of the subsequent analysis using correlation coefficient values and the MST graph model.
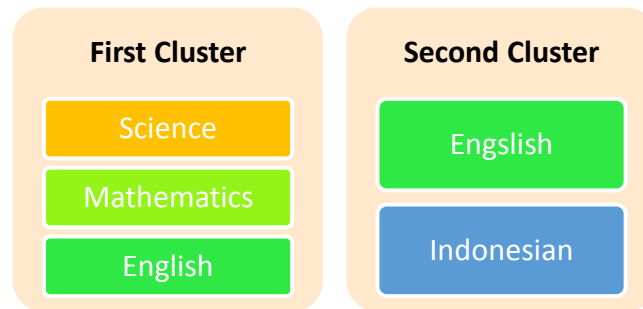
Figure 3. Initial Cluster of Subjects based on Scatterplot

Correlation Coefficient analysis

The correlation coefficient used in this paper is the Spearman correlation coefficient. Spearman correlation coefficient values can be used in groups of data that have a normal distribution or not, that is the reason for using these correlation coefficient values. The determination of the Spearman correlation coefficient is aided by SPSS software is in Table 3.

Table 3. Correlation Coefficient Values between Subjects

Variable	2017 Coeficient	2018 Coeficient
IND-ENG	0,646	0,696
IND-MATH	0,613	0,609
IND-SCIENCE	0,689	0,684
ENG-MATH	0,840	0,756
ENG-SCIENCE	0,847	0,756
MATH-SCIENCE	0,856	0,785

Table 3 shows some important information, three subjects; Mathematics, Science, and English show dominance as three subjects that have a close relationship. This reinforces the allegation in the previous analysis using scatterplot. As a result, the first cluster remains occupied by these three subjects. Slightly different from the previous second cluster. Table 3 shows that in 2017 and 2018 the relationship of Indonesian Language subjects and natural sciences was relatively consistent (strong) with scores of 0.689 and 0.684 respectively. While the relationship between Indonesian and English increased significantly in 2018 compared to 2017 with a correlation coefficient greater than Indonesian-IPA. The analysis provides a new cluster which is slightly different from the previous analysis with additional IPA in the second cluster, shown in Figure 4.
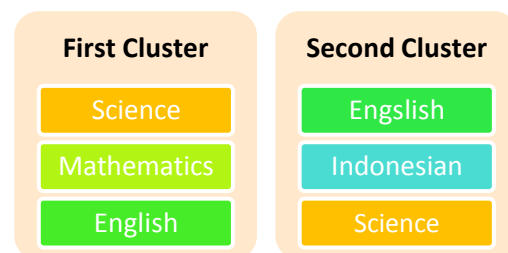
Figure 4. Second Cluster of Subjects based on Correlation Coefficient Values

The next analysis uses the MST graph model with the help of the Kruskal algorithm. The analysis will determine the final cluster on the results

of the national exams in Lampung Province in 2017 and 2018.

Maximum Spanning Tree Analysis

Based on the previous correlation coefficient, the initial graph model of the relationship between subjects in the national exams in Lampung province in 2017 and 2018 is shown in Figure 5.
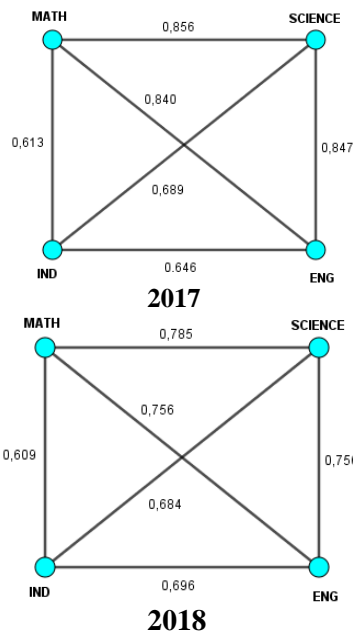

Figure 5. Initial Graph Model of Relationships between Subjects in 2017 and 2018.

The correlation coefficients between subjects listed in Table 3 are sorted from those that have the strongest to lowest relationships for each year, 2017 and 2018. The ordering aims to apply the Kruskal algorithm to form a maximum spanning tree (MST) graph. Table 4 is the correlation coefficient values that have been sorted for 2017. Table 5 is an orderly list of correlation coefficient values between subjects in the national exams in Lampung province in 2018 to form the second MST graph model.

Table 4. Sort List of Correlation Coefficient Values between Subjects in 2017

Weight	Edge
0,856	MATH-SCIENCE
0,847	ENG-SCIENCE
0,840	ENG-MATH
0,689	IND-SCIENCE
0,646	IND-ENG
0,613	IND-MATH

Table 5. Sort List of Correlation Coefficient Values between 2018 Subjects

Weight	Edge
0,785	MATH-SCIENCE
0,756	MATH-ENG
0,756	ENG-SCIENCE
0,696	IND-ENG
0,684	IND-SCIENCE
0,609	IND-MATH

The results of the national junior high school exams were eventually modeled by three different MST graphs, one 2017 data graph and two 2018 graphs as follows:
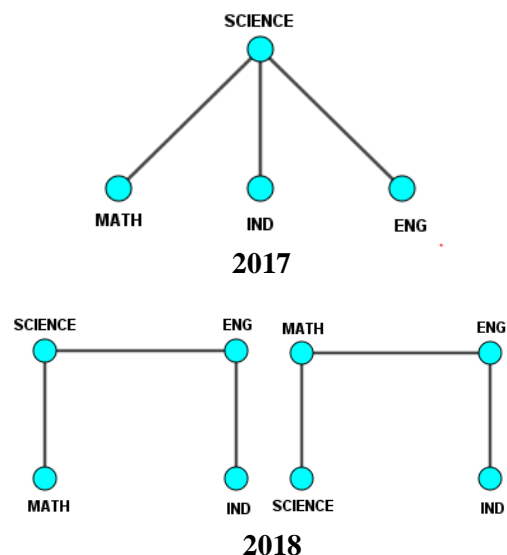

Figure 6. Graph of MST National Exam results in Lampung Province Junior High School in 2017 and 2018

Figure 6 shows the relationship of subjects in the Lampung province junior high school national examinations in 2017 and 2018 based on the principle of neighborhood on the MST graph. In the 2017 MST graph it is seen that absolute science subjects are the center of subjects compared to other subjects with the most branches or degrees of three, meaning that indirectly the ability of students in science lessons influences their learning outcomes in other subjects at the UN. Note that in the process of forming the 2017 MST graph, there is the MATH-SCIENCE-ENGLISH circuit. As a result, the closest subject relationship is owned by a group consisting of the three subjects previously mentioned (MATH-ENG-SCIENCE) and the subject that has the closest relationship with Indonesian is science with a correlation coefficient value of 0.684 so that subjects science is grouped with Indonesian.

Figure 6 on the 2018 MST graph model consists of two graph variations. For the first MST graph found on the left in figure 6, it has a center point in science and English subjects with the closest neighbors to science are mathematics and English, while English is next to science and Indonesian. Then, in the second MST graph (picture to the right), it also has two central points namely mathematics and English where mathematics is next to science and English while English is neighbor to mathematics and Indonesian. Based on the 2018 MST graph model, the first subject cluster is filled by SCIENCE-MATH-ENG and the second cluster is filled by SCIENCE-ENG-IND with the consideration that the relationship between mathematics and Indonesian is relatively more agile than with other subjects.

Based on the analysis of the MST graph model in the two years, the final clusters of subjects in the national exams in Lampung province based on 2017 and 2018 data are shown in Figure 7. Another advantage of clustering using the MST graph model is that it can sort the members of each cluster according to the most influence closely with other subjects.
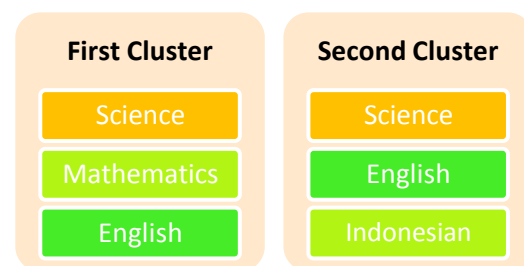

Figure 7. Final Clusters of Subjects for the National Examination Results in Lampung Province Junior High School

The first cluster does not change from scaterplot analysis, correlation coefficients, and was strengthened by analysis of the MST graph model. The final analysis reinforces the previous argument and adds information related to the order of the members of the subjects in the cluster according to the level of the subjects that most influences, so that the cluster members consist of Natural Sciences, Mathematics and English sequentially.

The members of the second cluster consist of science, English and Indonesian sequentially. The cluster was formed with the consideration that Science has a strong influence on all subjects in the Lampung provincial national exam based on the results of the 2017 MST graph. Therefore, it reinforces the second cluster argument based on the value of the correlation coefficient. In order, Science ranks first with consideration of the 2017 MST graph and other reasons, namely that

Indonesian only has a close relationship with English, so that the cluster members consist of IPA, English and Indonesian respectively.

Discussion

The last group formed shows that there is a close relationship in the subjects of Science, Mathematics and English in the first cluster and another relationship is formed by Science, English, and Indonesian in the second cluster. There are interesting things from the two clusters, pay attention to Figure 8, in the first cluster there is one language group subject, namely English, between Mathematics and Science and there is one science subject between two language subjects (English and Indonesian) in second cluster. In this section, a careful analysis of the two clusters is carried out through a theoretical study based on relevant sources.

The first cluster is the cluster that has the closest relationship based on the value of the correlation coefficient in Table 5. In this cluster there is a strong relationship between three subjects, English, Mathematics and Science. In this section, the interesting thing is the relationship between English and Mathematics - Science. In some places, in countries that do not use English as the main language, it shows that the influence of English language skills indirectly has a significant effect on academic achievement in Science and Mathematics of students in Grades 10, 11, and Higher Education at the beginning of the lecture year(Fenoll, 2018; Maskar, 2020; Murray, 2017; P. Ferrer & J. Dela Cruz, 2017; Stoffelsma & Spooren, 2019). The role of learning English in Mathematics is in the section on delivering the concept of Mathematics. Mathematical concepts consisting of definitions, theorems, postulates, and others come from English sources. Therefore, it is necessary to have good English skills in order to understand mathematical concepts well, and it will be easier to understand the arguments on the concept when referring to the main source directly. However, a significant effect on academic achievement can occur if the Mathematics Teacher understands mathematical concepts well and also has good English skills(Stacey, 2016). The 2017 and 2018 Lampung Provincial National Examination results show that this influence has not been able to support student academic achievement. This can be seen from the Mathematics average score on the UN score which is still below 50%, which correlates with the English score which also has an average value below 50%. Furthermore, the role of English in science learning is in the section on understanding the use of scientific terms. In addition, reading skills also have a major effect on increasing academic achievement in science learning(A. Imam, 2016; Hacioğlu et al., 2016). Reading skills in this cluster are certainly not reading skills directly in the learning process in class or taught by teachers, because the language of instruction in the majority of schools in Lampung Province uses Indonesian. Therefore, this reading skill is an additional part for students who like to observe and refer to a wider source, in English. This skill certainly needs to be supported by the financial strength of the student's parents, so that the influence of English reading skills on academic achievement in Science is influenced by other variables, such as facilities, money, home environment, and others(Prinsloo & Harvey, 2020). This influence is also still relatively low when referring to the 2017 and 2018

Lampung Province National Exam scores which have Science and English scores which are still below 50%. There needs to be a strengthening of English skills and abilities at the beginning of the year to support other subjects, especially Science and Mathematics.

To improve learning achievement in Mathematics through a combination of learning Science and English, contextual learning methods can be used. Contextual learning through a cultural approach is effective in improving students' understanding of mathematical concepts (Maskar, 2018; Maskar & Anderha, 2019; Wulantina & Maskar, 2019). The learning method through this cultural approach requires good skills in reading literacy and other relevant knowledge. Therefore, the right grouping of subjects is important so that learning objectives can be achieved. In addition, the use of technology also greatly supports students in understanding concepts and makes it easier to carry out learning simulations in class or when students are at home. The use of technology has a significant influence in helping teachers and students support the learning process. In learning Mathematics, GeoGebra is one of the technology-based media that can be used for free, but still has an optimal role(Maskar & Dewi, 2020; Maskar & Wulantina, 2019)

The second cluster consists of Indonesian, Science, and English. The discussion on the relationship between English and Science has been discussed previously. The relationship between English and Science lies in reading skills outside of the resources provided at school, so that students have more knowledge and can understand scientific terms in science learning through these skills. In this context, Indonesian language lessons should

have a more significant effect, considering that the general language of instruction in junior high schools in Lampung Province is Indonesian. In general, the correlation of Indonesian in science learning, which lies in students' reading skills so that they can understand concepts well. Meanwhile, the relationship between English and Science lies in the ability of students to understand scientific terms in Science Learning and the ability of students to be able to refer to wider sources. These skills certainly need to be optimized through the right techniques in science learning. This optimization can be done using a technique developed by Brown and Ryoo (Prinsloo & Harvey, 2020) in 2008 when they conducted research in Africa. The results show that the approach by introducing and teaching science phenomena uses everyday language and then switches to scientific language after that. The technique is called by teaching science concepts linguistically and ethnically. This method is significant in Africa because it can separate the concepts and linguistic components of Science. This technique is also a solution for minority learners or learners with difficult access to learning resources and lack of facilities in the school and home environment.

These clustering can use for determining the learning schedule, determining the number of lesson hours, determining the subjects that are prerequisites, and the group of subjects given every day, as well as other technical matters. In addition, this clustering is also useful for determining which subjects can be collaborated in order to run optimally. If there is a STEM or STEAM trend now, of course with a little modification, based on this reference, the collaboration can be

rearranged with optimal results because it is in accordance with the situation of students in the education unit.

However, referring to existing theoretical studies, the influence on each of these clusters is also influenced by other variables, one of the most important being the facilities and availability of learning resources for students, especially for students whose parents are in a weak economic level.

CONCLUSION AND SUGGESTION

There are two final clusters that were formed based on data on the Lampung Province National Examination (UN) scores in 2017 and 2018 with a total of 15,876 data consisting of the value of Science, Mathematics, English, and Indonesian subjects. The first cluster consists of Mathematics, Science, English, and the second cluster consists of Indonesian, English, and Science subjects.

The study on clustering is expected to be the initial basis for schools, especially in Lampung Province to be able to develop similar research with larger data so that the results can be used as a reference for decision making for technical and non-technical matters in order to improve student achievement in Lampung Province. Therefore, if this cluster is to be applied to policy making in schools or institutions where the majority are still lacking in terms of facilities and learning resources, there needs to be a solution for procuring these facilities first and then learning will run effectively and optimally by improving every aspect. technical management in education, improvement in competence in the curriculum, as well as increasing the competence of human resources, especially teachers.

The author hopes that this study will continue to be developed with greater data and at every level of education, in primary and secondary schools. Considering that the results of the National Examination produced more than 11 million data from the beginning of its implementation, of course it would be very unfortunate if the data was not processed into useful information for the improvement of education in Indonesia.

ACKNOWLEDGMENT

REFERENCES

A. Imam, O. (2016). Effects of Reading Skills on Students' Performance in Science and Mathematics in Public and Private Secondary Schools. *Journal of Education and Learning (EduLearn)*, *10*(2), 177–186. https://doi.org/10.11591/edulearn.v10i2.3430

Admin. (2017). *Tentang Ujian Nasional Berbasis Komputer*. Kemdikbud. https://unbk.kemdikbud.go.id/tentang

Cheng, D., Zhu, Q., Huang, J., Wu, Q., & Lijun, Y. (2014). Clustering with local density peaks-based minimum spanning tree. *IEEE Transactions on Knowledge and Data Engineering*, *13*(9), 1–15. https://doi.org/10.1109/tkde.2019.2930056

Crobe, A., Demuru, M., Didaci, L., Marcialis, G. L., & Fraschini, M. (2016). Minimum spanning tree and k -core decomposition as measure of subject-specific EEG traits. *Biomedical Physics &*

Engineering Express, *2*, 1–7. https://doi.org/10.1088/2057-1976/2/1/017001

Fenoll, A. A. (2018). English proficiency and mathematics test scores of immigrant children in the US. *Economics of Education Review*, *64*(April), 102–113. https://doi.org/10.1016/j.econedurev.2018.04.003

Hacıoğlu, Y., Yamak, H., & Kavak, N. (2016). Pre-service science teachers' cognitive structures regarding Science, Technology, Engineering, Mathematics (STEM) and science education. *Journal of Turkish Science Education*, *13*(Specialissue), 88–102. https://doi.org/10.12973/tused.10173a

Hussain, S., Atallah, R., & Kamsin, A. (2019). *Cybernetics and Algorithms in Intelligent Systems* (Vol. 765, Issue January). Springer International Publishing. https://doi.org/10.1007/978-3-319-91192-2

Kenneth H. Rosen. (2013). *Elementay Number Theory and Its Applications*. Pearson.

Kumar, B., & Pal, S. (2011). Mining Educational Data to Analyze Students Performance. *International Journal of Advanced Computer Science and Applications*, *2*(6). https://doi.org/10.14569/ijacsa.2011.020609

Li, Y., Gou, J., & Fan, Z. (2019). Educational data mining for students' performance based on fuzzy C-means clustering. *The Journal of Engineering*, *2019*(11), 8245–8250. https://doi.org/10.1049/joe.2019.0938

Lingaya, S. R., Gerardo, B. D., &

Medina, R. P. (2019). Modified graph-theoretic clustering algorithm for mining international linkages of philippine higher education institutions. *International Journal of Advanced Computer Science and Applications*, *10*(6), 90–95. https://doi.org/10.14569/ijacsa.2019.0100613

Maskar, S. (2018). Alternatif Penyusunan Materi Ekspresi Aljabar Untuk Siswa Smp/Mts Dengan Pendekatan Pendidikan. *Prisma*, *VII*(1), 53–69.

Maskar, S. (2020). Maximum Spanning Tree Graph Model: National Examination Data Analysis of Junior High School in Lampung Province. *Proceeding International Conference on Science and Engineering*, *3*(April), 375–378. https://doi.org/10.14421/icse.v3.531

Maskar, S., & Anderha, R. R. (2019). Pembelajaran transformasi geometri dengan pendekatan motif kain tapis lampung. *Mathema Journal*, *1*(1), 40–47.

Maskar, S., & Dewi, P. S. (2020). Praktikalitas dan Efektifitas Bahan Ajar Kalkulus Berbasis Daring Berbantuan Geogebra. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *4*(2), 888–899. https://doi.org/10.31004/cendekia.v4i2.326

Maskar, S., & Wulantina, E. (2019). Persepsi Peserta Didik terhadap Metode Blended Learning dengan Google Classroom. *Inomatika (Jurnal Pendidikan Matematika)*, *1*(2), 110–121. https://doi.org/https://doi.org/10.35438/inomatika.v1i2.156

Mishra, A., Bansal, R., & Singh, S. N. (2017). Educational data mining

and learning analysis. *Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering*, 491–494. https://doi.org/10.1109/CONFLUENCE.2017.7943201

Murray, M. (2017). How does the grade obtained at school for English and Mathematics affect the probability of graduation at a university? *Pythagoras*, *38*(1), 1–7. https://doi.org/10.4102/pythagoras.v38i1.335

P. Ferrer, F., & J. Dela Cruz, R. (2017). Correlation of Stem Students' Performance in the National Career Assessment Examination and Academic Subjects. *PEOPLE: International Journal of Social Sciences*, *3*(1), 532–541. https://doi.org/10.20319/pijss.2017.s31.532541

Prinsloo, C. H., & Harvey, J. C. (2020). The Differing Effect of Language Factors on Science and Mathematics Achievement Using TIMSS 2015 Data: South Africa. *Research in Science Education*, *50*(6), 2207–2226. https://doi.org/10.1007/s11165-018-9769-9

Shingari, I., Kumar, D., & Khetan, M. (2017). A review of applications of data mining techniques for prediction of students' performance in higher education. *Journal of Statistics and Management Systems*, *20*(4), 713–722. https://doi.org/10.1080/09720510.2017.1395191

Silva, C., & Fonseca, J. (2017). Educational Data Mining: A Literature Review. In *In Europe and MENA Cooperation Advances in Information and Communication Technologies*. Springer, Cham.

https://doi.org/10.1007/978-3-319-46568-5

Song, K., Li, C., Zhang, G., & Xing, C. (2019). Query based graph data reduction algorithms and application in education. *14th International Conference on Computer Science and Education, ICCSE 2019, Iccse*, 550–555. https://doi.org/10.1109/ICCSE.2019.8845357

Stacey, J. (2016). Does adding Mathematics to English language learners' timetables improve their acquisition of English? *Language Issues*, *27*(1), 84–87. http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=bri&AN=117114062&site=ehost-live

Stoffelsma, L., & Spooren, W. (2019). The Relationship Between English Reading Proficiency and Academic Achievement of First-Year Science and Mathematics Students in a Multilingual Context. *International Journal of Science and Mathematics Education*, *17*(5), 905–922. https://doi.org/10.1007/s10763-018-9905-z

Sugiyono. (2015). *Metode Penelitian Kombinasi (Mix Methods)*. Alfabeta.

Walter, N., & Dubey, S. K. (2016). Design and development of novice conceptual approach for minimum spanning tree. *Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering, Confluence 2016*, 713–717. https://doi.org/10.1109/CONFLUENCE.2016.7508212

Wattimena, A. Z., & Lawalatta, S. (2013). Aplikasi Algoritma Kruskal Dalam Pengotimalan Panjang Pipa. *BAREKENG: Jurnal*

Ilmu Matematika Dan Terapan, *7*(2), 13–18. https://doi.org/10.30598/barekengvol7iss2pp13-18

Wulantina, E., & Maskar, S. (2019). Pengembangan Bahan Ajar Matematika Berbasis Lampungnese Etnomatematics. *Edumatica: Jurnal Pendidikan Matematika*, *09*(02), 71–78. https://doi.org/https://doi.org/10.22437/edumatica.v9i02.7493

Yang, J. (2017). Clustering Analyzing of Undergraduate Schools based on K-Means Algorithm. *Proceedings of the 2017 International Conference on Advanced Mechatronic Systems*, 309–311.