STUDENT ACADEMIC STREAMING USING CLUSTERING TECHNIQUE

***Ely Salwana*[1]*, Suraya Hamid*[2]*, Norizan Mohd Yasin*[3]**

[1] Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Malaysia
[2, 3] Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.

E-mail: elysalwana@ukm.edu.my[1], suraya_hamid@um.edu.my[2], norizan@um.edu.my[3]

ABSTRACT

The balance of human capital supply and industry demands are crucial to sustain a competitive advantage in order to ensure stability of economic growth. Unfortunately, companies often find it hard to recruit the right people. Many ideas obviously relate to connect human capital to education because human capital is created through education. In the education sector, there is a critical need to develop an effective planning mechanism to distribute the students into the most suitable area in the industry. In order to achieve this, the student's pathway needs to be planned systematically in school by identifying the student's streaming based on their academic performance. In this case, students who have the same performance will be grouped in the same cluster using a data mining technique. However, the problem is that, it is difficult to identify real potential for the students, because their performance is not well monitored, and current assessment systems do not support the student's academic planning activities. Besides, there is no specific technique used to group students into clusters based on their performance. This study aims to overcome the problem by grouping the students according to their performance using clustering techniques and to propose a suitable model. This study aims to overcome the problem by identifying a suitable clustering model that can be used to analysis an educational data. The data involves is student performance data. Based on the data, two clusters of students are created which is science and arts. A novelty in the method of study is the use of three clustering models and a comparison among them in order to find a suitable clustering model to be used with student academic performance data. The study was conducted in five schools in Malaysia to support students' grouping in two different academic streams, which are science and art. The result demonstrated the best model of clustering technique that is suitable for mining the educational data. Moreover, suitable streaming based on the students' performance and education policy was created from the results. It can be used to assist schools and students in determining the appropriate streaming for the students, and support for the human capital needs by the country in the future.

Keywords: student's performance; data mining; clustering; academic streaming

1.0 INTRODUCTION

Few researchers such as Ndiyo[1], Kyriacon[2], Lan & Jamison[3] and Dasgupta & Weale[4] agree that education is the important element that will support economic growth of a country. Economic growth is supported by human capital needed by the industry[5]. The human capital that can be fitted into market needs is a valuable skilled person who can be produced by an excellent education system[6]. In education, it is important to identify and produce balanced manpower in two areas, which are science and the arts. In this case, there is a critical need to identify students' ability, so we can distribute students into groups according to their performance[7]. For the grouping process, data mining (DM) is used, as the data mining is a popular technique for clustering[7]–[13].

DM involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large datasets[14]. In the context of education, data mining is also known as educational data mining (EDM), which viewed a potential means to develop an algorithm for discovering the data that is derived from educational surroundings[15].

Educational data mining (EDM) is concerned with studying, building, and exploiting computerised techniques to identify patterns that involve big volumes of significant educational data that can be difficult and tough to explore[16]. EDM is formed from three main areas as shown in Figure 1; these are computer science, education, and statistics. The connection of these three areas also formulate other subareas that are linked to EDM such as computer based education, DM and machine learning, and learning analytics (LA)[15]. The EDM main task is to construct computational models to mine data that originated in an educational setting. Good practices in EDM can potentially answer important research questions about student performance.

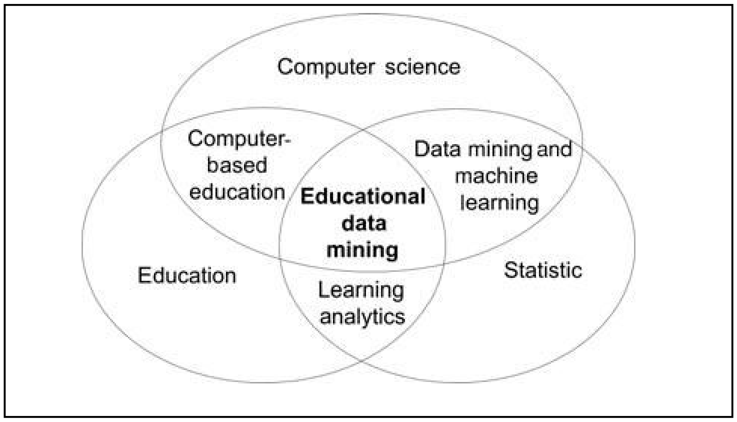


Fig. 1: Main Areas Related to Educational Data Mining [15]

The EDM that involves student performance data may introduce many significant effects to the education community specially to the student[12]. The grouping or clustering result can be used to support student academic planning, which is to guide the student's pathway. Therefore, this study aims to identify an appropriate data mining technique that can be used for mining an educational data, which is student performance data, in order to cluster students into groups. This clustering will follow education policy which is to identify balanced human capital needed by the industry that can be used as assets to support economic growth[17].

2.0 BACKGROUND

Generally, education contributes to the growth of an economy through the achievement of human capital on training and abilities. In the past few decades, there have been discussions and studies on the human capital involvement in the economic growth process[18]–[21].

A country relies on highly qualified human capital that is needed by the industry to develop their economies[17] and the human capital is produced through an excellent education system that is systematic and carefully planned[22]. Planning at school level, involves determining the direction or learning pathway for the students. Student Learning Pathway in education refers to academic streaming for a particular student[23], [24]. It can be described as a direction or pathway of the student through their development in school that will support and prepare them to meet and exceed their expectations[25]. Determination of academic streaming is done based on the students' academic performance for a specific subject.

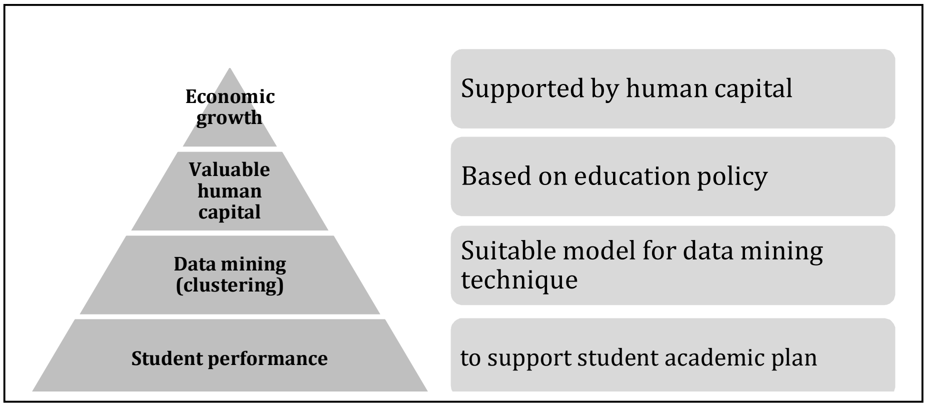

Economic growth
Supported by human capital
Valuable human capital
Based on education policy
Data mining (clustering)
Suitable model for data mining technique
Student performance
to support student academic plan

Fig. 2 : Students' Academic Performance for Economic Growth Using Data Mining

In this study, DM is proposed to cluster students based on their performance, so that we can identify the most suitable area for them. As mentioned in Section 1, DM techniques are widely used in education to analyze the student data. It helps to ensure that the knowledge that cannot be seen can be obtained from a large data. This is most related to student academic planning whereby it is a plan for setting up a student's pathway throughout the learning process in school[26]–[29]. The conceptual model of this study is illustrated in Figure 2.

From the previous studies [30]–[35] , the appropriate technique used to streaming students based on their performance is clustering. Clustering technique is used to group data of the same features and a popular technique for data mining [7]–[13]. Clustering technique discovers natural groupings of data sets or objects that have same features. This technique created several groups from the data that have the same characteristics. Using clustering data can be divided into natural groups and summarized the students' learning pattern and support for teaching and learning target [30].

Clustering has been used in schools to encounter the academic essentials of potential students [11]. Students with the same grade level are grouped together in one classroom [36]. Consequently, the students who are mathematically gifted will be placed in a classroom while the verbally gifted in another classroom.

Carol C. Burris & Allison [37] in their study for National Education Policy Centre (NEPC) explored the effect of sorting on overall student achievement or grouping and found a significant learning gains. This grouping technique involved a similar type of materials and assignments, teaching plan based on student needs, and interest-based instructional among others [38]. To sort the students, teachers and schools must use various sources of data to recognize their skills effectively. However, there are many unidentified students due to lack of resources [39].

Overall, cluster grouping can positively affect the performance of all students. A well-developed cluster grouping can offer a special program for high achieving students and attend a special need for those that are under performed. However, this grouping should be done dynamically based upon development needs and detailed learning condition [40]–[42].

In this study, the specific model of clustering technique is expected to group the students into two groups, which are science and art based on their academic performance. Only the science and art streams are considered in terms of the clustering process, as they are the main academic streaming in public schools in Malaysia. Normally in Malaysia, students will be grouped into these two streams. However, there are other specific academic streams available such as pure science, religion, religion as a profession, accounting, economics, and technical (drawing).

3.0 DATA COLLECTION

Data collection involved content analysis. The content analysis of documents used would be quantitative analysis. The documents used are students' academic performance data for year 2013. The content of the documents includes the academic performance of secondary three students from five schools under the Ministry of Education Malaysia (MOE), and the four main subjects (Malay, English, Mathematics and Science) that are used for identifying academic streaming.

In the data collection, the students' academic performance documents are used to run the analysis that is carried out for the purpose of discovering hidden patterns and relationships, as well as to identify the best natural cluster model for the educational data involved. This is because, based on the literature[8]–[10], [12], [43], [44], clustering is an important technique to support the process for students' academic streaming or distributing the students based on their performance. For this purpose, the method of analysis used is clustering analysis, using IBM SPSS Modeler® 15.0.

Table 1 : The Level / Grade of School in Malaysia [45]

Level/Grade	**Age**		
A. Preschool		C. Secondary school	
Pre-school playgroup	3-4	Secondary 1	13
Kindergarten	4-6	Secondary 2	14
B. Primary school		Secondary 3	15
Primary 1	7	Secondary 4	16
Primary 2	8	Secondary 5	17
Primary 3	9	D. Post-secondary education	
Primary 4	10	Tertiary education (College or University)	Ages vary
Primary 5	11		
Primary 6	12		

Table 2: Student's Academic Performance Data Used in the Clustering Analysis

Variable	**Description**	**Possible Values**
Id	Student id	Continuous number
Name	Student's name	String
bm (bm, bm2, bm3, bm4, bm5 – 5 times assessment in a year)	*Bahasa melayu(BM)* subject	1 (excellent) 2 ↑ 3 4 5 6 (poor)
bi (bi, bi2, bi3, bi4, bi5 – 5 times assessment in a year)	*English* (BI) subject	1 (excellent) 2 ↑ 3 4 5 6 (poor)
mt (mt, mt2, mt3, mt4, mt5 – 5 times assessment in a year)	Mathematics subject	1 (excellent) 2 ↑ 3 4 5 6 (poor)
Sc (sc, sc2, sc3, sc4, sc5 – 5 times assessment in a year)	Science subject	1 (excellent) 2 ↑ 3 4 5 6 (poor)

Class	Group for student whether in science class or art class	c1 = science group
		c2 = art group

In Malaysia, the school year is divided into two semesters. The first semester begins at the beginning of January and ends in May; the second semester begins in June and ends in November. The level or grade of school is divided into four stages, which are preschool, primary school, secondary school and post-secondary education. Table 1 shows level/grade of school in Malaysia. In secondary school levels there are two stages, which are lower secondary schools (secondary 1 – secondary 3) and high secondary schools (secondary 4 – secondary 5). During lower secondary level all students are taught the exactly same subjects. The academic streaming for placing students in science or art classes done after secondary three has been completed. That is the reason; this study used academic performance data from secondary three students.

The data was obtained from five schools throughout Malaysia, which involved 465 students. Table 2 describes the variables that were obtained from the documents used in the analysis.

Currently, MOE uses *Penilaian Berasaskan Sekolah* (PBS) or school-based assessment system. PBS is a holistic form of assessment that assesses cognitive (intellectual), affective (emotional and spiritual) and psychomotor (physical) skills in accordance with the National Education Philosophy and Curriculum Standard for Primary Schools (KSSR).

The assessment system uses band 1 to 6 to reflect student achievement as shown in Table 3.

Table 3: Band Using in Penilaian Berasaskan Sekolah (PBS)

Band	Standard
6	Know, understand and can do with exemplary ethics
5	Know, understand and can do with civilized admirable
4	Know, understand and can do with politeness
3	Know, understand and can do
2	Know and understand
1	Know

Band 1: Students know the basics, or can perform basic skills or respond to the basics.
Band 2: Students show understanding to change the form of communication or translate and explain what they have learned.
Band 3: Students can use the knowledge to perform their skills in a given situation.
Band 4: Students perform a skill politely, courteously or perform something systematically and follow the procedures.
Band 5: Students perform a skill politely in a new situation, systematically in accordance with the procedures, with a consistent and positive attitude.
Band 6: Students are able to express ideas that are creative and innovative and have the ability to make a decision in order to adapt everyday demands and challenges. They can also speak publicly using their own words in moral and consistently exemplary way, to get and disseminate information.

4.0 DATA ANALYSIS

Distributing students based on their data into natural clusters provides a good summary on how students are learning. It also benefits to school in order to plan about instructing and teaching [30]. Student clusters consist of groups of students who demonstrate similar learning curves throughout the whole year[43]. These clusters are helpful in identifying key activities that differentiate the students' performance.

For better decision making, large data require a proper method such as data mining (DM) to extract the knowledge from repositories. DM aims to find valuable information from large groups of data[46]. It focuses on using a few techniques and procedures in order to determine patterns of data[47].

The importance of DM is derived from model of technique development and problem analysis. After the problem is identified, it is used to assist in the creation of DM models. Alternative solutions are produced, and models are then developed to examine other alternatives. The best selection is then made and applied consistently with suitable models. There is no decision activity in this data mining process because typically, the activities join and blend together along the whole process, with regular looping back to previous stages to understand the problem.

Like other analytical research methods, content analysis requires that data to be examined and interpreted in order to elicit meaning, gain understanding, and develop empirical knowledge. In this study, the content analysis is used to identify academic streaming for the students. The documents involved were analyzed using analytical tool which is IBM SPSS Modeler® 15.0. The IBM SPSS Modeler® 15.0 is used in this research because it can discover patterns and trends in structured or unstructured data more easily, using a unique visual interface supported by advanced analytics [48]. It is also able to quickly build predictive models using educational knowledge and deploy them into daily operations to improve decision making [49].

In addition, SPSS Modeler offers a variety of modelling methods either using data from machine learning, artificial intelligence, or statistics. The methods allow us to derive new information from data and to develop predictive models. In addition, in this study it can be used to prepare data for analysis, find the best clustering model based on hidden patterns in data and quickly produce consistent and accurate results [50]. This is important to achieve one of the objectives of the study, which is to find the best clustering model that suits the educational data used. In this case, this study will be more focused on knowledge discovery rather than on technical tasks like scripting and coding.

This study suggests the use of three clustering models to identify suitable streaming for students in accordance with a students' performance in learning sessions throughout the year. The performance data were tested with IBM SPSS Modeler and three clustering models have been suggested based on the data used. In order to find the most appropriate clustering model, all these three models then be used to analysis the same data, and the result is compared.

The fundamental of any clustering algorithm is the measure of similarity of two patterns. First, student performance data for the whole year is collected. Next, instance and attribute-selection is used to transform into a numerical format for mining. Then, clustering is applied using SPSS Modeler to create the appropriate models and all these models are compared. This study used three models which are: K-means[51], TwoStep[52] and Kohonen[53]. The result of the analysis shows three important conclusions from the data, specifically; model summary, cluster size and predictor importance. From this information, an appropriate clustering model can be determined.

4.1 K-means

K-means is one of the simplest unsupervised learning models used for clustering[54]. It is the most widely-used, simple and well-known clustering algorithm, and in this research it targets to divide *n* instances into *k* clusters where by individual instance belongs to the cluster with the nearest mean[31] .

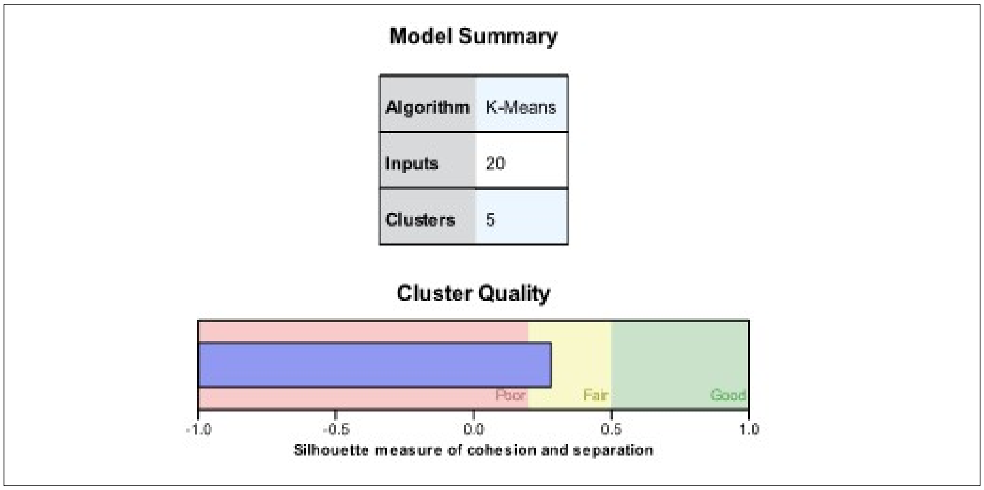
The K-means algorithm categorises data into a predefined number of clusters based on the Euclidean distance as the similarity measure[51]. Data of the particular cluster are connected with one centroid data, which represents the ''midpoint'' of the cluster. It is also the mean of the data that fit together. The key idea is to define *k* centroids, unique for each cluster. Different positions will cause different consequences, so all these centroids should be located in a smart way. The best way is to put them as far away from each other as possible. The next step is to proceeds individual point fit in to a specified dataset and links it to the closest centroid. When there is no pending point, the first step is to complete an early grouping. At this point, *k* new centroids must be re-calculated because midpoint of the clusters comes from the earlier step. Next, after the *k* new centroids are obtained, a loop has been generated in order to create new binding among the same dataset points and the closest new centroid. From this loop, it shows that *k* centroids adjust their position step by step up until no more adjustment happens or centroids do not change any more.

The main idea is:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right|^2$$

Where $\left\|x_i^{(j)} - c_j\right\|^2$ is a selected distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.
The algorithm is followed in the following steps:

1. Put k points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Allocate each object to the group that has the closest centroid.
3. When all objects have been allocated, recalculate the positions of the k centroids.
4. Repeat Steps 2 and 3 until the centroids no longer change. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.



Model Summary

Algorithm	K-Means
Inputs	20
Clusters	5

Fig. 4 : Cluster Size And The Cluster Quality Using K-Means Model

Using K-means, the first results to be presented are the cluster size and the cluster quality. Cluster size and cluster quality for this model are shown in Figure 4. Based on the analysis, cluster size is 5 and cluster quality is 0.282. The size of the smallest cluster using this model is 3, which is 4.5% of the overall cluster size. The size of the largest cluster is 23, which represents 34.3% of the overall cluster size. Meanwhile, the size ratio of the largest cluster to the smallest cluster is 7.67. The medium size of this ratio shows a modest cluster quality and the predictor importance is sc2 (which is the science subject for test No. 2).

4.2 TwoStep

This model combines the ability of the K-means clustering to handle a very large dataset, and the ability of the Hierarchical clustering (HCA – Hierarchical Cluster Analysis) to give a visual presentation of the results.

The TwoStep model is an exploratory tool designed to disclose natural groupings or clusters within a dataset that would otherwise not be apparent. The algorithm employed by this model has several desirable features that differentiate it from traditional clustering model [48]:

- **Handling of categorical and continuous variables.** By assuming variables to be independent, a joint multinomial-normal distribution can be placed on categorical and continuous variables.
- **Automatic selection of number of clusters.** By comparing the values of a model-choice criterion across different clustering solutions, the procedure can automatically determine the optimal number of clusters.
- **Scalability.** By constructing a cluster feature (CF) tree that summarises the records, the TwoStep algorithm allows large data files to be analysed.

The TwoStep clustering procedure consists of the following steps[50]:

1. Pre-clustering
   Pre-clusters are the clusters of original cases/objects that are used in place of raw data to reduce the size of the distance matrix between all possible pairs of cases. After completing the pre-clustering, the cases in the same pre-cluster are treated as a single entity[55]. Thus, the size of the distance matrix depends upon the number of pre-clusters instead of cases. Hierarchical clustering method is used on these pre-clusters instead of the original cases.

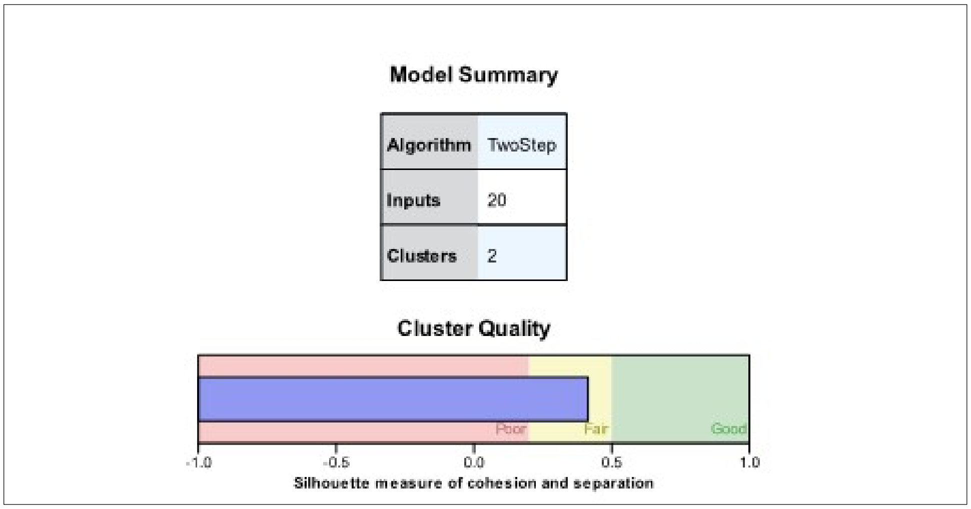
   The process is applied by building an adapted cluster feature (CF) tree. The CF tree contains levels of nodes, and each node comprises a number of items. A final sub-cluster is represents by a leaf item. A new record quickly guided into a correct leaf node by the non-leaf nodes and their items. Each item is categorized by its CF that contains the items' total of records, mean and variance of each range field. The items then counts for each group of each representative field. Beginning from the root node for each consecutive record, it is continually guided by the closest item in the node in order to discover the closest child node, and descends along the CF tree. When the process is getting a leaf node, it discovers the closest leaf item in the leaf node. The record is absorbed into the leaf item and the CF of that leaf entry is updated, if the record is inside a threshold distance of the closest leaf entry. Else, it starts its specific leaf item in the leaf node. If there is no space in the leaf node to produce a new leaf item, the leaf node is divided into two. The items in the original leaf node are separated into two groups using the farthest pair as seeds, and reallocating the remaining items based on the familiarity measure. If the CF tree develops beyond permitted maximum size, the CF tree is reconstructed based on the current CF tree by increasing the threshold distance criterion[56]. The reconstructed CF tree is reduced and later has space for new input records. This procedure remains until a whole data pass is completed.

   All records falling into the same item can be mutually characterized by the items' CF. Once a new record is added to an item, the new CF can be calculated from this new record and the previous CF without concerning about the specific records in the item. These assets of CF make it potential to retain only the item CFs, rather than the sets of specific records. Therefore the CF tree is much smaller than the original data and can be stored in memory more efficiently.

   The structure of the built CF tree may depend on the input order of the records. To reduce the order result, before constructing the model, the records are organized randomly.

2. Outlier handling
   A non-compulsory outlier-handling step is applied in the algorithm in the practice of constructing the CF tree. Outliers are reflected as data records that do not suitable into any cluster. Data records in a leaf item are reflected as outliers if the total of records is less than a definite portion (25% by default) of the size of the largest leaf item in the CF tree. The procedure search for possible outliers and puts them apart, before reconstructing the CF tree. After reconstructing the CF tree, the procedure checks whether these outliers suitable without increasing the tree size. Finally, small items that cannot fit in are outliers.

3. Clustering
   Sub-clusters, which is a non-outlier sub-clusters (if using outlier handling) from the pre-cluster step is taking as input and then groups them into the preferred number of clusters. Traditional clustering methods can be applied efficiently, as the total of sub-clusters is smaller than the total of original data. TwoStep uses an agglomerative hierarchical clustering method, since it works well with the auto-cluster method. Agglomerative hierarchical clustering is a process of merging clusters, only one cluster remains containing all records at the end of the process. The process begins by defining a preliminary cluster for individual sub-clusters that are produced in the pre-cluster step. All clusters produced are then compared, and the couple of clusters with the minimum distance among them is nominated and combined into a single cluster. The new set of clusters is compared, after the combination of the sub-clusters. The closest pair is combined, and the process repeats until all clusters have been combined. It is a similar process, on how a decision tree is built.



Model Summary

Algorithm	TwoStep
Inputs	20
Clusters	2

Fig. 5 : Cluster Size and The Cluster Quality Using Twostep Model

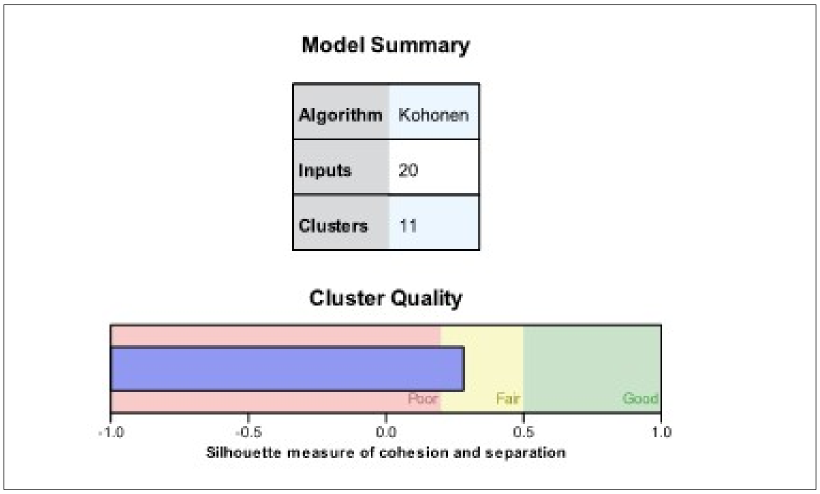
Based on the data, the first results from the TwoStep are the cluster size and the cluster quality. The cluster size and the cluster quality for this method are shown in Figure 5. Based on the analysis, the cluster size is 2 and the cluster quality is 0.413. The size of the smallest cluster using this model is 16, which is 23.9% of the overall cluster size, and the size of the largest cluster is 51, which represents 76.1% of the cluster size, whereas, the ratio of size of the largest cluster to the smallest cluster is 3.19. The small size of the ratio here showed a high cluster quality. For the predictor importance for the TwoStep, it shows relative importance of each field in estimating the model, the predictor importance measuring the *importance* of all variables. Normally, p*redictor* variables are correlated. Definition of relative *importance* of the *predictor* variables is a difficult procedure. However, SPSS Modeler can help to determine the predictor by selecting the model used. When applying TwoStep model, to the data the predictor importance is sc4 (science subject for test No. 4).

4.3 Kohonen

Kohonen is used because of its ability in mapping high-dimensional data to lower dimensions. It is applicable for exploratory data analysis. Its combination with U-Matrix method will help to detect data clusters much more easily than most classical methods.

The fundamental of Kohonen comes from knowledge in the brain. The brain uses internal space data image for keeping information. Firstly, data received are transformed to vectors which are encoded to the neural network[57], [58]. The output from such neural network is geometrically prepared to some arrangement, e.g. abreast, or to the rectangle and thus there is a possibility of neighbour identification. This layer is called Kohonen layer. Number of inputs entering to the network is equal to the input space dimension. Principle of cluster analysis is ability of the algorithm in the Kohonen model to set respective neurons to the clusters of submitted patterns and thus to distribute submitted patterns into the clusters.

Based on the data, the first results from the Kohonen model are the cluster size and the cluster quality. The cluster size and the cluster quality for this method are shown in Figure 6. Based on the analysis, the cluster size is 11 and the cluster quality is 0.273. The size of the smallest cluster using this model is 1 which is 1.5% of the overall cluster size, and size of the largest cluster shows a big difference which is 14 which represents 20.9% of the cluster size, whereas, the ratio of size of the largest cluster to the smallest cluster is 14.00. The big size of the ratio here, showed a low cluster quality. For the predictor importance using Kohonen, it shows relative importance of each field in estimating the model. Using Kohonen the predictor importance is sc5 (science subject for test no 5).



Model Summary

Algorithm	Kohonen
Inputs	20
Clusters	11

Fig. 6 : Cluster Size And The Cluster Quality Using Kohonen Model

5.0 RESULTS

Based on the conducted clustering analysis, the findings discover hidden patterns and relationships among the attributes in the educational data used. It was found that the most important predictor for all the three models involved (TwoStep, K-means and Kohonen) is the subject of science. This is because, from all the analysis that was done, it was found that the value of an important predictor for this subject is the highest when compared to the other subjects.

Besides, the analysis performed can also identify the best natural cluster model for the educational data. In this case, TwoStep provides a natural cluster that is suitable for the data, which is two clusters; because the data needs to be grouped into two categories, which are arts streaming or science streaming. Meanwhile, K-means and Kohonen provide a larger cluster, which is 5 clusters for K-means and 11 clusters for Kohonen.

Table 4 : Summary for The Three Models Used in The Analysis

Clustering Algorithm	Cluster Size	Clustering Quality	Importance	Ratio of Size
TwoStep	2	0.413	0.807	3.19
Kmeans	5	0.282	0.496	7.67
Kohonen	11	0.273	0.478	14.00

All the three models used in the analysis are summarised, in terms of cluster size, clustering quality, the importance of the cluster and ratio size of the cluster. The cluster quality of TwoStep is 0.413, which is close to fair quality, but the quality of K-means and Kohonen are fairly low at 0.282 (K-means) and 0.273 (Kohonen). The scale used is between -1 and 1. A value of 1 indicates good quality, whereas, a value of -1 indicates poor quality. A value of −1 would mean all cases are located in the cluster centres of some other cluster. Due to the number of clusters produced by the K-means and Kohonen, the ratio of the cluster size is also high. This is because the calculation of the ratio of cluster size is made based on the largest cluster divided by the smallest cluster. The small ratio shows no significant difference in the distribution of the clusters. Among the three models analysed, TwoStep shows the smallest value of ratio of size which is 3.19, while for K-means and Kohonen the ratio of size is 7.67 and 14.00, the smaller size of ratio indicating better quality of clusters produced.

Based on the analysis, Table 4 shows that TwoStep is the suitable model that suits the educational data of students' performance, where the importance value is 0.807 compared to K-means (0.496) and Kohonen (0.478). This shows that based on the student's performance, TwoStep is the suitable model to group students into the two categories or academic streaming classes of art or science.

6.0 CONCLUSION

This study made a contribution to the existing knowledge on the academic streaming that is carried out in schools and the technology that can successfully be used to achieve a more systematic implementation.

The current study considered the clustering technique, which has been discussed in the literature. Taking into account the specific model in clustering techniques that are suitable for educational data, steps were taken to reach a decision on how best to use clustering to improve academic streaming. Cluster sampling is a relatively straightforward technique when there are already natural clusters or groupings within a population, which are expected to affect the outcome of data analysis. As the basis of this study is to define groups, which in this case are proportions of students that will be selected to undertake a particular academic pathway, clustering is a suitable technique. The selection of a commonly used model makes the selection process more transparent, and there is clarity about how the model has been developed and on which data the model has been based.

The TwoStep clustering model was identified as the most suitable, as it involves both K-Means and Hierarchical Cluster Analysis. The data was embedded to produce natural clusters that could be used to support the distribution of students across the two main streams of science or arts. The modelling experiment aimed to group the students into two clusters, based on their academic performance, which could potentially be used as a basis for streaming. The clustering method would be a basis by which students would also be able to identify their academic strengths and which stream would be more suitable for them when starting secondary level education. The TwoStep model was considered to be suitable for the purposes of this study on academic performance data, as it supports the development of natural clusters based on student academic performance data, which would be relevant to adhere to the Ministry of Education (MOE) guidelines on the 60:40 ratio of student assigned to science or arts-based streams.

Besides the development of an academic streaming model and the proposed clustering method, the same data from this study can be used to provide valuable information for the future development of students and schools. It will also help students to gain a better understanding about suitable career fields for their future. Furthermore, it can enhance the teaching and learning process because teachers will be informed about the academic performance of individual students and their learning pathway, and this information will help them to put every effort into developing these students. A well-developed cluster grouping can place students in the appropriate academic streams, and help teachers to better fulfil the needs of students because by placing high achievers in one classroom, their chances of having their needs met will increase, while other students will have the opportunity to develop in the other classrooms. This is vital for generating skills in specific fields to achieve the government's plan for meeting the human capital needs of the country.

For the future research, an extension of the scope of studies on clustering can be consider, in order to include other academic streams, apart from science and arts, that are available in government schools such as pure science, religion, religion as a profession, accounting, economics, and technical (drawing). When all existing streams are taking into account, students will put extra focus in identifying streams that are appropriate with their performance. Other than that, the overall contribution of the research could have been improved with an iterative feedback mechanism whereby the outcome of the clustering technique can be verified against the empirical study from the current practices and the improvements can be determined or measured.

REFERENCES

[1] N. Ndiyo, "The Paradox of Education and Economic Growth in Nigeria: An Empirical Evidence," in *NES Proceedings*, 2002.

[2] G. Kyriacon, "Level and Growth Effects of Human Capital: A Cross Country Study of the Convergence Hypothesis," New York, 1980.

[3] L. Lan and L. Jamison, "Impact of Education by Region," 1991.

[4] P. Dasgupta and M. Weale, "On measuring the quality of life," *World Dev.*, vol. 20, no. 1, pp. 119–131,

Jan. 1992.

[5] A. Boughanmi, “Human capital and economic growth,” *Bus. Rev. Cambridge*, vol. 13, no. 2, pp. 252–260, 2009.

[6] D. Olaniyan and T. Okemakinde, “Human capital theory: implications for educational development,” *Pakistan J. Soc. Sci.*, vol. 24, no. 2, pp. 157–162, 2008.

[7] E. Osmanbegović and M. Suljić, “Data Mining Approach for Predicting Student Performance,” *Econ. Rev. (Journal Econ. Business)*, vol. X, no. 1, pp. 3–12, 2012.

[8] M. Lopez and J. Luna, “Classification via Clustering for Predicting Final Marks Based on Student Participation in Forums,” in *Proceedings of the 5th international conference on educational data mining*, 2012, pp. 148–151.

[9] O. Oyelade, O. Oladipupo, and I. Obagbuwa, “Application of k means clustering algorithm for prediction of students academic performance,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 7, no. 1, pp. 292–295, 2010.

[10] M. Shovon, H. Islam, and M. Haque, “An Approach of Improving Students Academic Performance by using k means clustering algorithm and Decision tree,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 8, pp. 146–149, 2012.

[11] A. Zimmermann and L. Raedt, “Cluster-grouping: from subgroup discovery to clustering,” *Mach. Learn.*, vol. 77, no. 1, pp. 125–159, Jun. 2009.

[12] M. M. A. Tair and A. M. El-halees, “Mining Educational Data to Improve Students’ Performance : A Case Study,” *Int. J. Inf. Commun. Technol. Res.*, vol. 2, no. 2, pp. 140–146, 2012.

[13] F. a. Bachtiar, C. W. Eric, and K. Kamei, “Student grouping by neural network based on affective factors in learning English,” *Proceeding Int. Conf. e-Education, Entertain. e-Management*, pp. 226–229, Dec. 2011.

[14] J. Seifert, “Data mining: An overview,” *Natl. Secur. issues*, 2004.

[15] C. Romero and S. Ventura, “Data mining in education,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 12–27, Jan. 2013.

[16] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura, “Predicting students’ final performance from participation in on-line discussion forums,” *Comput. Educ.*, vol. 68, pp. 458–472, Oct. 2013.

[17] F. Schütt, “The Importance of Human Capital for Economic Growth,” *Inst. World Econ. Int. Manag.*, 2003.

[18] P. Romer, “Endogenous Technological Change,” *J. Polit. Econ.*, vol. 98, no. 5, 1990.

[19] P. M. Romer, “Increasing Returns and Long-Run Growth,” *J. Polit. Econ.*, vol. 94, no. 5, pp. 1002–1037, 1986.

[20] R. E. Lucas, “On the mechanics of economic development,” *J. Monet. Econ.*, vol. 22, p. 42, 1988.

[21] J. U. Umo, *Moving Nigeria Into the New Economy: A Human Capital Perspective*. Nigeria: Nigerian Economic Society, 2007.

[22] E. Hanushek and L. Wößmann, “The role of education quality for economic growth,” *World Bank Policy Res. Work. …*, 2007.

[23] C. Cooper, C. T. G. Coll, W. T. Bartko, H. M. Davis, and C. Chatman, *Developmental Pathways*

Through Middle Childhood: Rethinking Contexts and Diversity as Resources. Psychology Press, 2006.

[24] A. Taylor, "Pathways for Youth to the Labour Market : An Overview of High School Initiatives," Ottawa, Ontario, Canada, Ontario, Canada, 2007.

[25] K. Mittendorff, W. Jochems, F. Meijers, and P. den Brok, "Differences and similarities in the use of the portfolio and personal development plan for career guidance in various vocational schools in The Netherlands," *J. Vocat. Educ. Train.*, vol. 60, no. 1, pp. 75–91, Mar. 2008.

[26] J. S. Eccles, M. N. Vida, and B. Barber, "The Relation of Early Adolescents' College Plans and Both Academic Ability and Task-Value Beliefs to Subsequent College Enrollment," *J. Early Adolesc.*, vol. 24, no. 1, pp. 63–77, Feb. 2004.

[27] P. Barrett, Y. Zhang, J. Moffat, and K. Kobbacy, "A holistic, multi-level analysis identifying the impact of classroom design on pupils' learning," *Build. Environ.*, vol. 59, pp. 678–689, Jan. 2013.

[28] M. J. Gilbert, "The Relationship Between Pupil Control Ideology and Academic Optimism," Seton Hall University, 2012.

[29] B. Schneider, J. Judy, C. M. Ebmeye, and M. Broda, "Trust in Elementary and Secondary Urban Schools: A Pathway for Student Success and College Ambition," in *Trust and School Life*, D. Van Maele, P. B. Forsyth, and M. Van Houtte, Eds. Dordrecht: Springer Netherlands, 2014, pp. 37–47.

[30] W. Hämäläinen, V. Kumpulainen, and M. Mozgovoy, "Evaluation of clustering methods for adaptive learning systems," in *Artificial Intelligence Applications in Distance Education*, 2013, pp. 1–32.

[31] B. Firouzi, M. Sadeghi, and T. Niknam, "A new hybrid algorithm based on PSO, SA, and K-means for cluster analysis," *Int. J. Innov. Comput. Inf. Control*, vol. 6, no. 7, pp. 3177–4198, 2010.

[32] Y.-T. Kao, E. Zahara, and I.-W. Kao, "A hybridized approach to data clustering," *Expert Syst. Appl.*, vol. 34, no. 3, pp. 1754–1762, Apr. 2008.

[33] M. Laszlo and S. Mukherjee, "A genetic algorithm that exchanges neighboring centers for k-means clustering," *Pattern Recognit. Lett.*, vol. 28, no. 16, pp. 2359–2366, Dec. 2007.

[34] T. Niknam and B. Amiri, "An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis," *Appl. Soft Comput.*, vol. 10, no. 1, pp. 183–197, Jan. 2010.

[35] M. Yin, Y. Hu, F. Yang, X. Li, and W. Gu, "A novel hybrid K-harmonic means and gravitational search algorithm approach for clustering," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9319–9324, Aug. 2011.

[36] M. L. Gentry, "Promoting Student Achievement and Exemplary Classroom Practices Through Cluster Grouping: A Research-Based Alternative to Heterogeneous Elementary Classrooms," United States, 1999.

[37] C. C. Burris and K. E. Allison, "Review of Does Sorting Students Improve Test Scores?," Boulder, CO, 2013.

[38] D. W. Caldwell, "Educating Gifted Students in the Regular Classroom : Efficacy , Attitudes , and Differentiation of Instruction," Georgia Southern University, 2012.

[39] Jennifer Stepanek, "Meeting the Needs of Gifted Students : Differentiating Mathematics and Science Instruction," 1999.

[40] N. Bellin, O. Dunge, and C. Gunzenhauser, "The importance of class composition for reading achievement : Migration background , social composition , and instructional practices An analysis of the German 2006 PIRLS data," *IERI Monogr. Ser. Issues Methodol. Large-Scale Assessments*, vol. 3, no. 1,

pp. 9–34, 2006.

[41] L. Benson, "Serving gifted students through inclusion: A teacher's perspective," *Roeper Rev.*, vol. 24, no. 3, pp. 126–127, Mar. 2002.

[42] R. Burns and D. Mason, "Class composition and student achievement in elementary schools," *Am. Educ. Res. J.*, vol. 39, no. 1, pp. 207–233, 2002.

[43] H. Bian, "Clustering Student Learning Activity Data.," in *EDM*, 2010, pp. 277–278.

[44] P. Ajith, M. Sai, and B. Tejaswi, "Evaluation of Student Performance: An Outlier Detection Perspective," *Int. J. Innov. Technol. Explor. Eng.*, vol. 2, no. 2, pp. 40–44, 2013.

[45] Ministry of Education Malaysia, "The Public Schooling System - for Primary, Secondary and Post-secondary Levels," in *Schools of Malaysia Directory*, 4th ed., Kuala Lumpur: Challenger Concept (M) Sdn Bhd, 2013.

[46] B. K. Baradwaj, "Mining Educational Data to Analyze Students' Performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 6, pp. 63–69, 2011.

[47] P. Shirwaikar and N. Rajadhyax, "Data Mining on Educational Domain," *arXiv Prepr. arXiv1207.1535*, pp. 1–6, 2012.

[48] IBM Corporation, "IBM SPSS Modeler Text Analytics 15 User ' s Guide," United States, 2012.

[49] B. R. Devi, K. N. Rao, S. P. Setty, and M. N. Rao, "Disaster Prediction System Using IBM SPSS Data Mining Tool," *Int. J. Eng. Trends Technol.*, vol. 4, no. August, pp. 3352–3357, 2013.

[50] IBM Corporation, "IBM SPSS Modeler Professional," United States, 2011.

[51] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*. New York: John Wiley & Sons, 1990.

[52] M. Shih, J. Jheng, and L. Lai, "A two-step method for clustering mixed categroical and numeric data," *Tamkang J. Sci. Eng.*, vol. 13, no. 1, pp. 11–19, 2010.

[53] W. Snyder, D. Nissman, V. Bout, and G. Bilbro, "Kohonen Networks and Clustering: Comparative Performance of Colour Clustering," *Adv. Neural Inf. Process.*, pp. 984–990, 1991.

[54] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[55] J. P. Verma, *Data Analysis in Management with SPSS Software*. India: Springer India, 2013.

[56] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," in *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1996, vol. 1, pp. 103–114.

[57] S. Kajan, I. Sekaj, and M. Lajtman, "Cluster Analysis Aplications in Matlab Using Kohonen Network," in *Technical Computing Prague, 19th Annual Conference Proceeding*, 2011, vol. 2, no. 1.

[58] IBM Corporation, *IBM SPSS Modeler 15 Modeling Nodes*. 2012.