JOURNAL OF COMPUTATIONAL INNOVATION AND ANALYTICS
https://e-journal.uum.edu.my/index.php/jcia

How to cite this article:
Alawi, S. J. S., Shaharanee, I. N. M., & Jamil, J. M. (2023). Clustering student performance data using k-means algorithms. *Journal of Computational Innovation and Analytics, 2(1),* 41-55. https://doi.org/10.32890/jcia2023.2.1.3

CLUSTERING STUDENT PERFORMANCE DATA USING k-MEANS ALGORITHMS

[1]Sultan Juma Sultan Alawi, [2]Izwan Nizal Mohd Shaharanee & [3]Jastini Mohd Jamil
[1]Ministry of Education, South Alsharqia, Sultanate of Oman
[1,2&3]School of Quantitative Sciences,
Universiti Utara Malaysia, Malaysia

[3]Corresponding author: jastini@uum.edu.my

Received: 13/7/2022 Revised: 20/9/2022 Accepted: 26/10/2022 Published: 30/1/2022

ABSTRACT

Education institutions store large amounts of data regarding students, such as demographics, academic-related data, and student activities. These data were recorded and stored in many ways, including different filing systems and database formats. By having these data, education institutions have a better way to manage and understand their students. In addition, information related to their students can easily be accessed and extracted. As more data is recorded and stored, this could allow the educational institution to make more informed decisions and give educators good insight into the educational system. The research approach known as educational data mining (EDM) focuses on using data mining techniques to extract massive data from the educational context and transform it into knowledge that can improve educational systems and decisions. Clustering, an unsupervised learning technique, is one of the most powerful machine-learning tools for discovering patterns and unseen data. This work aims to provide insights into the data obtained from Oman Education Portal (OEP) related to the student's performance by manipulating the k-means algorithm.

Keywords: Clustering, education data mining, k-means algorithm, student performance.

INTRODUCTION

Nowadays, education may stimulate economic growth, increase gender equality, and reduce poverty. It can also improve health. Investing in education will pay off for many generations. Other than that, the capacity to influence society and the economy lies in the education sector, where it plays a vital role in the development of human capital and is linked with an opportunity for better living and improving life quality. For educators, the standard of student performance continues to be paramount. It is meant to make a difference on various levels: locally, regionally, nationally, and globally. Educators and researchers worldwide are interested in exploring the factors contributing to students' performance. Traditionally, student performance is assessed using tests and examinations. Tests and examinations based on past questions from the specific tests can give an approximate estimate of how well the student might perform in the actual test. However, nowadays, with the advancement of technology, many platforms gather a huge amount of data regarding students, such as information on their enrollment, collaboration, behavior in school, as well as interaction with colleges and teachers (Beal et al., 2006). Apart from that, this data can be utilized to understand factors that impact student performance (Li & Yoo, 2006). The Ministry of Education in the Sultanate of Oman established a new platform called Oman Education Portal (OEP) in 2007 (International Labour Organization, 2012). The main objective of the OEP is to offer useful services for students, teachers, parents, and decision-makers. For example, this may include capturing student life cycles, such as student personal information, e-books, exams, attendance, interactive classroom, results, and certificates. Note that this platform's student transactions and records generate huge data. The challenge that confronts the decision-makers in the Ministry of Education in the Sultanate of

Oman is how one can analyze this big data to improve the education system of the country as a whole.

Currently, educators in the Sultanate of Oman still rely on classic statistical methods to analyze big data generated in OEP information (Alawi et al., 2017). In certain situations, this method cannot be utilized, especially for the huge data collected from the OEP portal. The current approach to analyzing the data lacks many issues owing to the existence of a large number of records and missing and imbalanced data. As the data size increases significantly, many traditional methods that worked successfully in the past fail to operate effectively today. The huge amount of data generated in the OEP portal can provide clearer and deeper insights into students' behavior and future achievement. With a better understanding of students at risk of failing in their educational process and detecting the factors that impact their learning, educational institutions will be able to introduce strategies to improve student performance (Iam-On & Boongoen, 2017).

In recent years, education data mining (EDM) has emerged as one of the main methods of discovering useful and meaningful information. EDM is a strategy and a process of digging into massive data to extract useful information and the meaning of data. In addition, the EDM methods predict future trends and behaviors, which can be useful for decision-makers to make active, knowledgeable data-driven decisions. Association rule mining, classification, and clustering include several data mining techniques that are popular in the EDM community (Nandakumar, 2018).

This research paper focuses on utilizing a clustering algorithm to profile and group student performance characteristics obtained from the OEP dataset. Subsequently, this research aims to gain insight into how clustering analysis can be employed in the educational domain, which offers a new way to identify different characteristics of the student based on the clustering algorithms.

OMAN EDUCATION PORTAL (OEP) DATABASE

Currently, the Ministry of Education in the Sultanate of Oman manages 11 educational zones, 1,647 schools, 67,901 teachers and administrators, as well as 724,395 students (National Centre for

Statistics and Information, 2015). In addition, exchanging knowledge, training, thoughts, experiences, and skills in each institution is complicated due to geographic distance. Due to these challenges and issues, the Ministry of Education established a new educational platform called OEP in 2007 (International Labour Organization, 2012).

The objective of this platform is to provide good services for all the stakeholders in the ministry. For example, it can help parents manage their child's enrollment, control taken courses, and inquire about the student's academic and non-academic performance. The OEP also offers students access to school timetables, exam timetables, evaluation reports, and virtual classrooms. For teachers, it allows them to communicate with students through e-learning, follow up on students' achievements, and allocate students to activities and students' reports. For other administrators in the ministry, it allows them to perform supervision, training, authorization for related services, and e-correspondence.

All these services generate a huge amount of data for the Ministry of Education. These data are a valuable asset for the decision makers that can help to put effective strategies and plans for the Ministry of Education. It will help to improve the education system to be more efficient. However, a big dataset needs a special approach for data processing and an instrument that can convert massive data to support decision-makers in data interpretation and analysis.

EDUCATION DATA MINING (EDM)

The process of analyzing data from various angles and distilling it into usable information is known as data mining. The EDM, a sub-domain of data mining, is a trending discipline that focuses on applying various methodologies, tools, and algorithms in the exploratory, graphical and intelligent analysis of educational data repositories. It aims to discover and extract new structures, which in turn helps to understand, predict and improve the students' academic performances (Kumar & Radhika, 2014).

One of the important applications of EDM is discovering and understanding student performance. This includes identifying students' characteristics for improving their academic performances. Managing and improving student performance is crucial in this process. It offers a way for educational institutions to detect and distinguish students with high performance or, on the contrary, a student at risk of failing. The EDM aims to understand how students learn and identify various aspects that can improve learning and other educational activities. Moreover, the EDM processes provide real-time feedback exchange or improve learning management, enhancing the students' learning processes. The EDM provides a conduit for lecturers/teachers/instructors to investigate, monitor, and take student-centered actions to improve the students' learning process. EDM can be categorized into prediction, relationship mining, and clustering tasks.

Most experts agree that these categories of EDM techniques are common to different types of data mining (Darcan & Badur, 2012). The prediction usually works to develop a model which can infer a single aspect of the data (e.g., predicted variable) from some combination of other aspects of data (e.g., predictor variables). The second category is relationship mining, where the main objective is to discover relationships between the variables in a dataset with many variables. This may determine which variables are the most strongly associated with a single variable of particular interest. The last is clustering, a process of grouping objects into classes based on similarity. It is an unsupervised partitioning or classification of patterns (observations) into groups (clusters) based on their neighborhood within N-dimensional space. Clustering is particularly useful in cases where the most common categories within the data set are not known in advance.

This paper focuses on applying the clustering algorithm, which is the most suitable for detecting student performance in the OEP dataset. The objective of this research is to gain insight into how clustering analysis can be done in the educational domain and to highlight the potential characteristics of the students, which can be discovered using clustering algorithms.

K-MEANS CLUSTERING ALGORITHM

An unsupervised machine learning technique called clustering divides the input dataset into groups of objects more similar than those in other groups. There are various clustering techniques for developing useful clusters and segments. One of the crucial topics for data mining applications is clustering, which organizes items into groups based on their characteristics so that those in the same group are similar and those in different groups are distinct (Trivedi et al., 2011). The main advantage of clustering is that hidden patterns and structures can be detected by analyzing large datasets with little or no background knowledge. The clustering technique is a widely used technique for the future prediction of students' academic performance. Therefore, the k-means is the most popular and well-known clustering method (Patil & Baidari, 2019). Among various clustering techniques, a number of studies used the k-means algorithm because of its ease of use, simplicity, and performance (Nelson, 2015).

There are three major techniques in the clustering approach: partitioning, hierarchical, and density-based (Alashwal et al., 2019). The partitioning method can be used to determine the k clusters, optimizing each cluster based on the distance function. Hierarchical methods create a homogenous group by breaking a database into smaller groups. Meanwhile, a density-based method is used for low-density noise regions to find the dense regions available in the data space to separate one data group from another (Bernard, 2015) we first survey the research done on clustering analysis in education and identify the algorithms used. We then present a case-based experiment to show the relative performance of clustering algorithms with Learning Management System log data. We compare partition-based (K-Means.

The k-means algorithm is one of the existing methods of partitioning clustering. It is an evolutionary algorithm that gains its name from its method of operation, in which the algorithm clusters observations into the k groups, where k is provided as an input parameter (Bernard, 2015)we first survey the research done on clustering analysis in education and identify the algorithms used. We then present a case-based experiment to show the relative performance of clustering algorithms with Learning Management System log data. We compare partition-based (K-Means. Subsequently, it assigns each observation to clusters based on the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed, and the process begins again iteratively. The algorithm includes the following steps:

1. The algorithm randomly picks *k* points as the original cluster centers ("means").

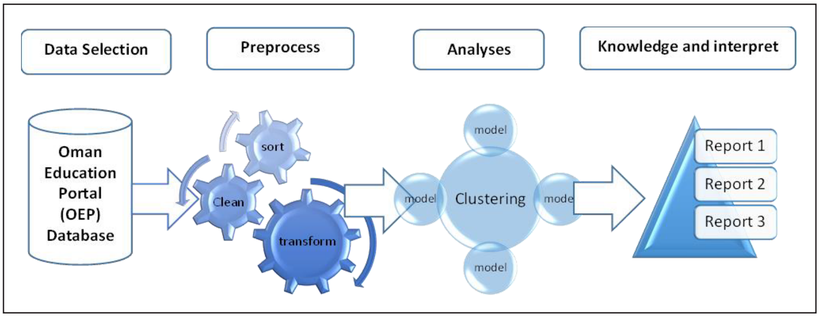
2. Each point in the dataset is assigned to the closed cluster based on the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeats until the clusters converge. Convergence means if the output of repeating Steps 2 and 3 does not make a material difference in the definition of clusters or no observations change in clusters.

METHODOLOGY

The methodological workflow of this work, as depicted in Figure 1, proceeds in the following steps: data selection, pre-processing, k-means cluster analysis, and data extraction. The pre-processing involves data categorization and exploratory analysis.

Figure 1

Research Methodology



Data Selection

Data analyzed in this study was extracted from tracking the data captured from the OEP database in the academic year 2019/2018. The OEP includes a large database that contains administrator data, teacher data, parent data, schedule data, etc. In this study, we are extracting only student data related to our project goal. The student dataset contains 49,588 student records regarding demographic information and each student's academic performance. Note that the demographic information includes the region, age, gender, religion, nationality, school size, class size, school time work, and final marks at the end of the year. In addition, all students recorded in grade 8 show how they studied in the academic year 2019/2018. The student record was collected from the OEP data warehouse.

Data Pre-processing

Several data pre-processing techniques were employed in this stage to obtain a suitable dataset for clustering development. The first technique is removing irrelevant attributes. The OEP database contained a large amount of data. Thus, only suitable and relevant attributes were considered in this analysis. For example, we used "studentID" to identify students and omitted other attributes like "First Name," "Second Name," and "Family Name." The second technique is the data transformation technique. This approach converted numeric to a categorical variable. For example, the age attribute (continuous) is transformed into three groups (under-class age, in-class age, above-class age). For the nationality variable, we changed the data into two groups (Omani and non-Omani). Meanwhile, for the father's education level , we used four groups (non-educated, read & write, school level, and higher education). For the size of the school and class, we utilized three groups (low, medium, and high). The teacher-to-student ratio group was split into three groups: low, medium, and high. The final variable included in the clustering model is student performance. We grouped all the grades into four possible values: Excellent (from 1100 to 900), Good (from 900 to 650), Average (from 650 to 450), and Failure (under 450), which is shown in Table 1.

Table 1

Variables Description

Variable name	Values	Variable Description
Region	Urban, rural	The location of the living student
School shift	Morning, evening	Schoolwork time in the morning or evening
Nationality	Omani, non-Omani	Student's nationality

(continued)

Variable name	Values	Variable Description
Gender	Male, female	Student's gender
Religion	Muslim, non-Muslim	Student's religion
Age	Under-class age, in-class age, above -class age	Students' age in a group
Father's education level	Non-educated, read and write only, school level, higher education	Student father's qualification certification
Class size	Low, medium, high	Number of students in the class
School size	Low, medium, high	Number of students in school
Teachers' ratio	Low, medium, high	Student/teacher ratio
Student performance	Excellent, good, average, failure	Final marks

The exploratory analysis gives a clear understanding of how the attributes are distributed in the student dataset. For instance, students' categorization based on the region is 80.7% from a rural area, 19.3% from an urban area, and 98.8% students studying in the morning to 1.2% in the evening. In a group of variables showing student nationality, there were 96.4% Omani and 3.6% non-Omani. In the gender variable, we found 50.7% males and 49.3 females. Other than that, a group of attendance included three categorizations with 97.1% normal absence, 1.6 medium, and 1.3 high absence. A group of variables based on the age of the students demonstrated 87.7% in the class age and 12.2% above the class age. In a class size group that presents several students in each class, we found 74.8% of students in the medium class, 23.9% in the high class, and 1.3% in the low class, respectively. In a group of school size, which presents a number of students in school, we found 50.9% in medium school, 39% high, and 10.1% in low school. In a group of teachers ratio, we present the number of students tutored by each teacher, 59.3% in the medium group, 34.5% in the high group, and 6.2% in the low group. The last variable is the target one, demonstrating student performance, which we divided into four groups. Here, we discovered 40.1% excellent, 52.8% good, 4.9% average, and 2.2% failure students.

Table 2

Data Distribution and Characteristics

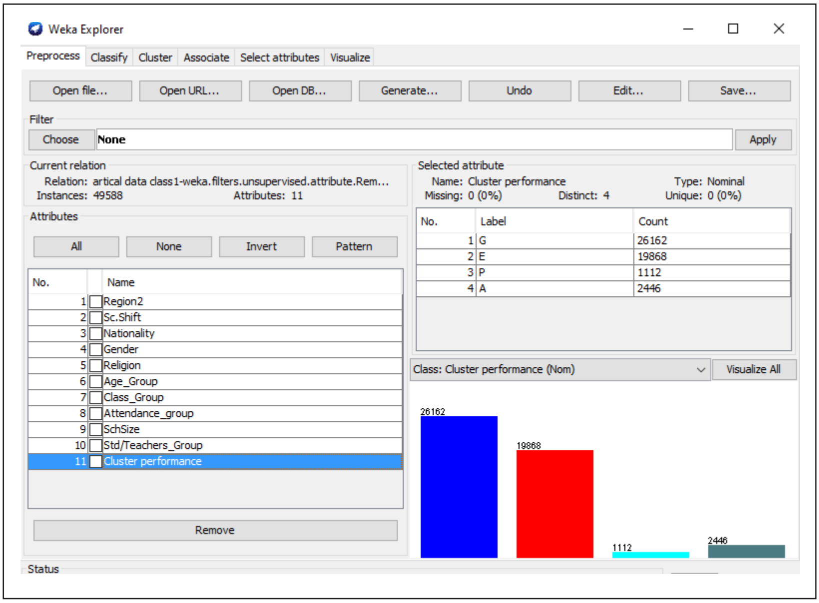
Variables	Values	Students' Dataset	
		Number of Instances	Proportion (%)
Region	Rural	40028	80.7
	Urban	9560	19.3
School shift	Evening	573	1.2
	Morning	49015	98.8
Nationality	Omani	47790	96.4
	Non-Omani	1798	3.6
Gender	Female	24423	49.3
	Male	25165	50.7
Attendance	High	621	1.3
	Medium	794	1.6
	Normal	48173	97.1
Age	Under class age	19	0.0
	In class age	43495	87.7
	Above class age	6074	12.2
Class size	High	11846	23.9
	Low	665	1.3
	Medium	37077	74.8
School size	High	19346	39.0
	Low	4988	10.1
	Medium	25254	50.9
Teachers ratio	High	17104	34.5
	Low	3054	6.2
	Medium	29430	59.3
Performance	Excellent	19868	40.1
	Good	26162	52.8
	Average	2446	4.9
	Failure	1112	2.2

RESULTS

This research demonstrated the applicability and effective use of the k-means algorithm with Weka software. Figure 2 depicts the OEP data loaded into the Weka software. As depicted in Figure 2, there are 49,588 student records with 11 attributes as variables, with the target cluster groups being divided into four performance groups. The "Excellent" group contained 19868 records, the "Good" group contained 26162 records, the "Average" group contained 2446 records, and the "Failure" group contained 1,112 records.

Figure 2

Data Loading in Weka Software



Filter: None

Current relation
Relation: artical data class1-weka.filters.unsupervised.attribute.Rem...
Instances: 49588 Attributes: 11

Selected attribute
Name: Cluster performance Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count
1	G	26162
2	E	19868
3	P	1112
4	A	2446

Attributes

No.	Name
1	□ Region2
2	□ Sc.Shift
3	□ Nationality
4	□ Gender
5	□ Religion
6	□ Age_Group
7	□ Class_Group
8	□ Attendance_group
9	□ SchSize
10	□ Std/Teachers_Group
11	□ Cluster performance

Class: Cluster performance (Nom)

Clustering Results

In this research work, four clusters were pre-selected as this can mimic the real proportions of the student performance in a dataset. Tables 3, 4, and 5 present the traits of the four clusters.

Table 3

Cluster Characteristics

Attribute	Cluster 0 (4621)	Cluster 1 (27509)	Cluster 2 (16527)	Cluster 3 (931)
Gender	Male	Female	Male	Female
Age	In class age	In class age	In class age	Above class age
Region	Rural	Rural	Rural	Rural
School shift	Morning	Morning	Morning	Morning
Nationality	Omani	Omani	Omani	Omani
Religion	Muslim	Muslim	Muslim	Muslim
Class size	Medium	Medium	Medium	Medium
School size	Low	Medium	High	Medium
Attendance	Normal	Normal	Normal	Normal
Teacher ratio	Low	Medium	High	High

The cluster is divided into four categories: Excellent, Good, Average, and Failure. The distribution of student performance in the clusters is shown in Table 4.

Table 4

Distribution of the Instances

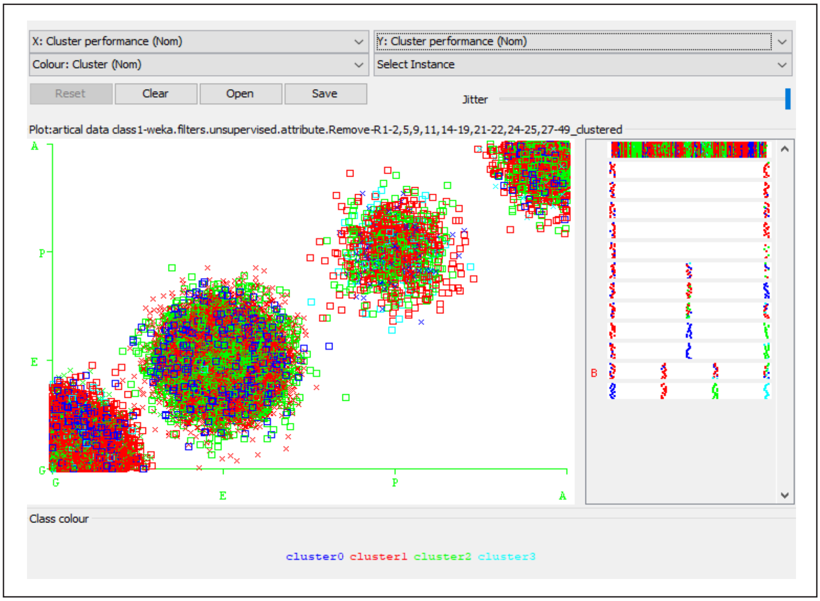
	Instances	%	Cluster description
Cluster 0	4621	9%	Failure
Cluster 1	27509	55%	Excellent
Cluster 2	16527	33%	Good
Cluster 3	931	2%	Average

In comparison with other groups, Cluster 1 is profiled as an excellent student. This group holds nearly 55% of the total number of students, 27509. The majority of students from this group were female students of class age studying in a medium school class size. Meanwhile, Cluster 2 is a group of good-performance students. This group has the second-highest number of students at 16,527 (33%). However, this group has a huge school size of students. Overall, we can conclude that a typical student in this group has a good performance level. On the other hand, Cluster 3 is a group of students with average performance.

This group consists of 931 students and shows the lowest number of students. Nevertheless, the students in this group demonstrated a fair level of performance. Finally, Cluster 0 is a group of weak-performance students, with the number of students in this group being 4621 (9%). In this group, most of the students are males. We observed that this group includes the small school size of students with weak performance. Figure 3 depicts the visual output of Weka, where the instances can be represented in a two-dimensional graph. One of the interesting outputs from this analysis is that gender and age of the student play an important role in identifying student performance. It is worth mentioning that female students have higher performance than male students, while the failed student primarily represents male students in the above-age class.

Figure 3

Clusters of Student Performance



X: Cluster performance (Nom)

Y: Cluster performance (Nom)

Colour: Cluster (Nom)

Plot:artical data class1-weka.filters.unsupervised.attribute.Remove-R1-2,5,9,11,14-19,21-22,24-25,27-49_clustered

CONCLUSION

In this paper, four clusters of student performance were identified based on their characteristics and school performance based on the data obtained from the OEP portal. This cluster can help educators, especially from the Ministry of Education in Oman, to identify students with the highest risk of failing and underperforming. Furthermore, this approach can help academic planners track students' progress during their study curriculum. Consequently, this model is crucial to assist academic planners in evaluating students and understanding the causes of the drop in students' performance and their related characteristics. The study contributed to the applications of IT methods and statistical data analysis to social and education studies. A special case was using a clustering approach to Omani students and schoolchildren.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., & Moustafa, A. A. (2019). The application of unsupervised clustering methods to Alzheimer's disease. *Frontiers in Computational Neuroscience*, *13*(May), 1–9. http://doi.org/10.3389/fncom.2019.00031

Alawi, S. J. S., Shaharanee, I. N. M., & Jamil, J. M. (2017). Profiling Oman education data using data mining approach. *AIP Conference Proceedings*, *1891*. http://doi.org/10.1063/1.5005467

Beal, C. R., Qu, L., & Lee, H. (2006). Classifying learner engagement through integration of multiple data sources. *Proceedings of the National Conference on Artificial Intelligence*, *21*(1), 151. http://www.aaai.org/Papers/AAAI/2006/AAAI06-024.pdf

Bernard, K. D. M. (2015). Comparative performance analysis of clustering techniques in educational. *IADIS International Journal on Computer Science and Information Systems*, *10*(2), 65–78.

Darcan, O., & Badur, B. (2012). Student profiling on academic performance using cluster analysis. *Journal of E-Learning & Higher Education*, *2012*, 1–8. http://doi.org/10.5171/2012.622480

Iam-On, N., & Boongoen, T. (2017). Generating descriptive model for student dropout: a review of clustering approach. *Human-Centric Computing and Information Sciences*, *7*(1), 1–24. http://doi.org/10.1186/s13673-016-0083-0

International Labour Organization. (2012). *Sultanate of Oman - Extension of the Decent Work Country Program*. https://www.ilo.org/wcmsp5/groups/public/---arabstates/---ro-beirut/documents/genericdocument/wcms_446085.pdf

Kumar, A. D., & Radhika, V. (2014). A survey on predicting student performance. *International Journal of Computer Science and Information Technologies*, *5*(5), 6147–6149.

Li, C., & Yoo, J. (2006). Modeling student online learning using clustering. *Proceedings of the 44th Annual Southeast Regional Conference,* 186-191.https://doi.org/10.1145/1185448.1185490

Nandakumar, A. N. (2018). An effective cure clustering algorithm in education data mining techniques to valuate student's performance. *International Journal of Applied Engineering Research, 13*(10), 7493–7498.

National Centre for Statistics and Information. (2015). His Majesty Sultan Qaboos bin Said. *Statical Year Book*, (42).

Nelson, K. (2015). Using k -means clustering to model students LMS participation in traditional courses. *Issues in Information Systems, 16(4)*, *16*(Iv), 102–110.

Patil, C., & Baidari, I. (2019). Estimating the optimal number of clusters k in a dataset using data depth. *Data Science and Engineering*, *4*(2), 132–140. http://doi.org/10.1007/s41019-019-0091-y

Trivedi, S., Pardos, Z., Sárközy, G., & Heffernan, N. (2011). Spectral clustering in educational data mining. *Proceedings of the 4th International Conference on Educational Data Mining*, 129–138.