Creative Research Publishers
https://www.crossref.org/06members/50go-live.html

Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 8, No 1, 2022 Series

Data Mining Technique for Improved Academic Performance

[1]Onwuama T.U., [2]Okpalla C.L. & [3]Enyindah P.
[1&2]Dept. of Computer Science, Federal University of Technology, Owerri, Nigeria
[3]Dept. of Computer Science, University of Port Harcourt, Port Harcourt, Rivers State, Nigeria
E-mails: udora09@gmail.com, lilymmao@yahoo.com, promise.enyindah@uniport.edu.ng

ABSTRACT

Data and information are very important in any system or organization, and their appropriate utilization can generate breakthrough in any field. In the academic field, their importance cannot be over emphasized. This paper applies the clustering data mining technique to detect various patterns that form students' characteristics in order to predict, and analyze students' academic performance. The students' data from previous semester were collected coupled with other individual socio-cultural factors, and the data was mined using the Simple K Means Clustering algorithm. In the results, it was observed that not only does the academic performance of students depend on the class test grade or exam performance, but contributing influential factors such as family, background, attitude to success or failure, commitment to learning, class lectures etc., can significantly affect academic student's performance. Our results show a high efficiency of our system. This study will help teachers to improve students' academic performance.

Keywords: Data Mining, Educational Mining, K-means Clustering, Algorithm, Association Rules.

Journal Reference Format:
Onwuama T.U., Okpalla C.L. & Enyindah P. (2022): Data Mining Technique for Improved Academic Performance, Digital Humanities and Development Research. Vol. 8.No. 1, Pp 23-34. Available online at https://www.isteams.net/behavioralinformaticsjournal
DOI No - dx.doi.org/10.22624/AIMS/BHI/V8N1P3

1. INTRODUCTION

According to Brojo et al., (2016), data mining also known as knowledge discovery is a process through which huge databases can be identified of various novel, valid and recognizable patterns which are hidden. Data mining discloses the hidden information through the rapidly growing data piles by using a set of operations. The evolution of data mining began when data was first stored in books, journals, notepads, diaries etc. This made it possible for useful information to be accessed, retrieved or amended. This evolution continued when data was later stored on computers, coupled with improvements in data access, and technologies that allow users to navigate through their data in real time. In recent years, there has been an increase in the need for data and information to be written, stored and useful pattern of information to be retrieved and easily forecasted for purposeful reasons ranging from educational down to commercial. The use of data mining to ask, determine and discover answers to questions within educational research is termed Educational Data Mining. Educational Data Mining methods often differ from methods from the broader data mining concept. It exploits the multiple levels of meaningful hierarchy in educational data. These methods (psychometrics) are often integrated to achieve this goal. The knowledge here is hidden among the educational data sets and it is extractable through data mining techniques.

Knowledge discovered by the teacher through traditional methods such as; choice of subject/course enrolled by the students, question and answer sessions, classroom teaching models, detection of abnormal values in the result sheets of the students, punctuality to classes etc., provide findings to help teachers understand their students' learning performance within a moderate class setting. With this, appropriate counsel was then provided by the teacher. Similarly, the concrete impacts of fairly rare individual differences have been difficult to statistically study with traditional methods. Thus, educational data mining has the potential to extend a much wider tool set to the analysis of important questions in individual differences. Just as a human teacher can adapt to an individual student, the same teacher can also learn more about how students learn, reflect and improve their practice by studying a group of students through data mining methods in education.

While a large focus of research in this area is to provide adaptation to a learner using the data stored in his/her student model, this paper is aimed at exploring ways to improve students' performance using data mining methods with traditional methods to draw conclusions on the capabilities of students to progress in their chosen fields of learning.

2. RELATED LITERATURE

2.1 Concept of Data Mining

The advent of information technology in various fields has led to large volumes of data storage in various formats like records, files, documents, images, sound, videos, scientific data and many new data formats. The data collected from different applications require proper methods of retrieving knowledge from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data. Data mining is a powerful technology which is helping enterprises and institutions of learning to turn data and information into knowledge. It is a Knowledge Discovery Process (KDP) that seeks to discover consistent patterns and/or systematic relationships between variables and then to validate the findings by applying the detected patterns to new subsets of data. Pattern identification and data processing to discover trends within the information is what data mining is all about.

Data Mining became prevalent with the advent of big data. Big data caused an explosion with the use of extensive data set records using data mining techniques to get relatively simple and straight forward statistics out of the system. There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns developing methods that discover knowledge from data originating from educational environments (Han and Kamber, 2000). Data Mining when used in education, especially to support reflection on teaching and learning systems is addressed as Educational Data Mining (EDM).

Educational data mining is the use of multiple analytical techniques to better understand relationships, structure, patterns, and causal pathways within complex datasets in institutions of learning. Educational Data Mining refers to tools, methods, techniques, and research designed for automatically extracting meaning from large repositories of data generated by or related to people's learning activities in educational settings. This field seeks to develop and improve methods for exploring this data in order to discover new insights about how people learn in the context of such setting. Learning Analytics is a closely related context with more emphasis on getting concise information about automatically getting data, along with human observation of the teaching and learning contexts in education settings. In summary, Educational Data Mining focuses on developing new tools and algorithms for discovering data patterns. Learning Analytics focuses on applying tools and techniques at larger scales in

instructional systems. Educational systems are increasingly engineered to capture and store data on user's interaction with the system. These data (whether big data, system log data, trace data etc.) can be analyzed using statistical machine learning and data mining techniques.
It uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor, and many others.

2.2 History of Educational Data Mining

The analysis of educational data is not a new development. Recent advances in educational technology including the increase in computing power and the ability to select, retrieve and store fine-grained data about students' use of a computer-based learning environment, have led to an increased interest in developing techniques for analyzing the large amounts of data generated in educational settings. This interest has led to a series of Educational Data Mining workshops held as part of several international research conferences such as:

1) In 2008, a group of researchers established what has become an annual international research conference on Educational Data Mining, the first of which took place in Montreal, Canada.
2) In 2009, the Journal of Educational Data Mining, an academic journal for sharing and disseminating research results.
3) In 2011, Educational Data Mining researchers established the International Educational Data Mining Society to connect EDM researchers and continue to grow the field.

Also, the introduction of public educational data repositories in 2008, such as the Pittsburgh Science of Learning Centre (PSLC) Data shop and the National Centre for Educational Statistics (NCES), public data sets have made educational data mining more accessible and feasible, contributing to its growth.

2.3 Goals of Educational Data Mining

A. Predicting Students' Future Learning Behaviour

This goal can be achieved with the use of student modeling by creating student models that incorporate the learner's characteristics, including detailed information such as their knowledge, behaviors and motivation to learn. Also, the Student's (user) experience of the learner and their overall satisfaction with learning are also measured.

B. Discovering or Improving Domain Models

Through the various methods and applications of EDM, discovering new models or improvements to existing models is possible. It involves finding or developing existing models that characterize the subject matter to be learned (for example; Math, Science etc.). Also identifying fruitful advanced teaching and learning sequences, and suggest how these sequences might be adapted to students' needs.

C. Studying the effects of Educational Support

This can be achieved through learning systems by studying the effects of varied learning process enhancements on student learning.

D. Advancing Scientific Knowledge about learning and learners through building models of learning processes that incorporate data about students, teachers, understanding of subject matter and principles from learning the subjects.

E. Supporting Learning for All Students by adapting learning resources to fit the particular needs identified, including adaptations for individual students when warranted.

2.4 Users and Stakeholders in Educational Data Mining

There are four main users and stakeholders involved with educational data mining. These include;

i. Learners
ii. Educators
iii. Researchers
iv. Administrators

2.4.1 Learners

Learners are interested in understanding students' needs and methods to the learner's experience and performance. For example, learners can also benefit from the discovered knowledge by using the EDM tools to suggest activities and resources that they can use based on their interactions with the online learning tool and insights from past or similar learners. For young learners, educational data mining can also inform parents about their child's learning progress.

2.4.2 Educators

Educators attempt to understand the learning process and the methods they can use to improve their teaching methods. They can use the applications of EDM to determine how to organize and structure the curriculum, best methods to deliver course information and the tools to use to engage their learners for optimal learning outcomes.

2.4.3 Researchers

These users focus on the development and the evaluation of data mining techniques for effectiveness. The wide range of topics in EDM ranges from using data mining to improve institutional effectiveness to student performance.

2.4.4 Administrators

These users are responsible for allocating the resources for implementation in institutions. As institutions are increasingly held responsible for student success, the administering of EDM applications is becoming more common in educational settings. Faculty and advisors are becoming more proactive in identifying and addressing at-risk students. It is sometimes difficult to get the information to the decision makers to administer the application in a timely and efficient manner.

2.5 Data Mining Algorithms

The most commonly used of these algorithms as mentioned by Harun et al., (2017) are as follows:

1. Association Rules
2. Clustering
3. Classification
4. Regression

Brojo et al., (2016) further added

5. Anomaly detection and
6. Summarization

Association rules: It is a data mining rule that is trying to find the relationship of common patterns of the objects within a data set. For example, which products are sold together? Which DNA profiles are sensitive to new drug? With this algorithm questions are proposed (Harun et al 2017). Brojo et al., (2016) extended the description of association rules by saying that in this model significant dependencies between variables are defined. Though it is a very simple method to be used but it is capable of providing a lot of insight and information related to the day to day business. This information can be used to generate the required revenue and even improve the efficiency of the business.

There are far fledged applications related to this method which can help various industries and business to increase their value. Here are some examples: Up-selling and cross-selling of products, physical organization of items, network analysis, and marketing and management. This method was used for many years in the industry for the market basket analysis but now new recommendations have been made by the engineers, which have overpowered the traditional methods.

Clustering Algorithm: As explained by Brojo et al., (2016), clustering is an important technique through which object grouping can be done (like the different groups of customers). The objects belonging to the same cluster are similar but those which are in the different groups are different. In this descriptive task a finite set of clusters are determined which identify or describe the data. The process of clustering can be defined in such a way that if you have a group of data points which have attributes of their own and have some kind of similarity then they should be clustered in such a way that the data points in that cluster are much alike each other. Data points in separate clusters are likely to be dissimilar to one another. To find how close or far one cluster is from the another, we can use the Euclidean distance, which can be applied only if attributes are continuous or other similarity measures that is relevant to the specific problem.

A useful application of clustering is marketing segmentation, in which distinct set of customers are made in the market and distinct marketing strategies are applied to each of the subsets. It is possible to do this by analysing the lifestyle related and geographical information of each customer and make their clusters. This will help in finding out the clustering quality of the customers by observing the difference in the buying patterns of the customers in one cluster to the customers in the other cluster. (Harun et al., 2017) said that clustering is used for grouping of similar objects in database. For example, determination of different customer profile in markets, while making city planning determination of house prices according to the geographical grouping.

Classification Data Mining Algorithm: Identifies which class the incoming information belong according to Harun et al., (2017). Brojo et al., (2016) explained that before digging into the hectic modelling phase of the analysis of data the primary step we have to take is classification. This classifies the data item in anyone of the predefined classes. Assume you have a set of records which have their own set of attributes and one of the present attribute is our class (as per the letter grades).

Our main motive is to find a model for the class that will be able to predict the undiscovered records (from external similar data sources) accurately which will be similar to the known label of the class, provided all values of other attributes. We usually divide the data set into two subsets, to train the model in a particular manner for a specific task: training set and test set. The model will be built with the help of training set and the test set will do the validation. It is the test set which determines the accuracy and performance of the model.

Regression: It is the data mining algorithm which determines relationship between more than one variable. As an example can be given as the determination of expenses of potential customers whose income and profession are known (Harun et al., 2017). In their work, Brojo et al., (2016) said that regression can be simply called as the "predictive power". Assuming a linear or non linear model of dependency, regression analysis can be used by us to predict the value of given (continuous) features based on the other features in the data. The data item is mapped into a real valued prediction variable. Here are some examples: The revenue of new products are predicted depending upon the complementary products.

Based on the amount of food and cigarette consumed by a person and his age the prediction of cancer can be done. "Logistic regression" is such a term which appears in almost every aspect of this field and regression techniques are also found to be useful in this science. These techniques are especially used in the case of neural network which can be used to create such complex functions which help in imitating the functionalities of the brain.

Anomaly Detection (change and deviation detection): Brojo et al., (2016) explained that anomaly detection technique helps to determine the most significant data change that has taken place in the database. This is calculated and identified on the previously determined data.

Summarization: With the help of summarization, a subset of the data present in the database is evaluated and a consequently a compact description is found (Brojo et al., (2016).

2.6 Related Works

Pandey and Pal (2011) conducted a study on the performance of students based by selecting 600 students from different colleges in certain universities in India. By means of Bayes Classification on category, language and background qualification, it was to find out whether new comer students will perform or not. The study had it that students from a stable home and have a good understanding of the English language performed better than the others.

Hijazi and Naqvi (2006) conducted a study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis that was stated as "Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' family income, students' mother's age and mother's education are significantly related with student performance" was framed. By means of simple linear regression analysis, it was found that the factors like mother's education and student's family income were highly correlated with the student academic performance.

Khan (2005), conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Galit (2007), gave a case study that use students' data to analyze their learning behavior to predict the results and to warn students at risk before their final exams. Al-Radaideh (2006), applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the Naïve Bayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models.

3. DATA MINING ALGORITHMS

3.1 Simple K- means Clustering

A predictive model known as Simple K-means Clustering was developed to predict the students' performance in an academic session. The datasets used for stimulation were taken partly from an online Harvard-MIT Person-Course Dataset and partly modified by the authors. In Cluster Analysis, k-mean algorithm is a method which has a goal of partitioning n-observations into k clusters in which each observation belongs to the cluster with the nearest mean. Clustering is one of the basic techniques often used in analyzing data sets and considered the most important unsupervised learning technique. This study makes use of cluster analysis to segment students into groups according to their characteristics, and then to group the student according to their behaviour during lectures and then it is mined together with other factors in order to determine the students' performance in an academic session. In these steps, k-means clustering algorithm was applied to the proposed data to get or mine valuable information from the dataset.

3.2 K-means Algorithm

In cluster analysis, k-mean algorithm is a method which has a goal of partitioning *n* observations into *k* clusters in which each observations belong to the cluster with the nearest mean. It is one of the simplest unsupervised learning algorithms that solves the well- known clustering problem. K-means algorithm initializes the cluster means by randomly generating *k* points in the data space. This is typically done by generating a value uniformly at random, within the range for each dimension.

The k-means algorithm follows the following steps:

i. Place *k* points into the space represented by the objects that are being clustered. These points represent initial group centroids.
ii. Assign each object to the group that has the closest centroid.
iii. When all objects have been assigned, recalculate the positions of the *k* centroids.
iv. Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated. The flowchart of the model is shown in Figure 1

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 8, No 1, 2022 Series

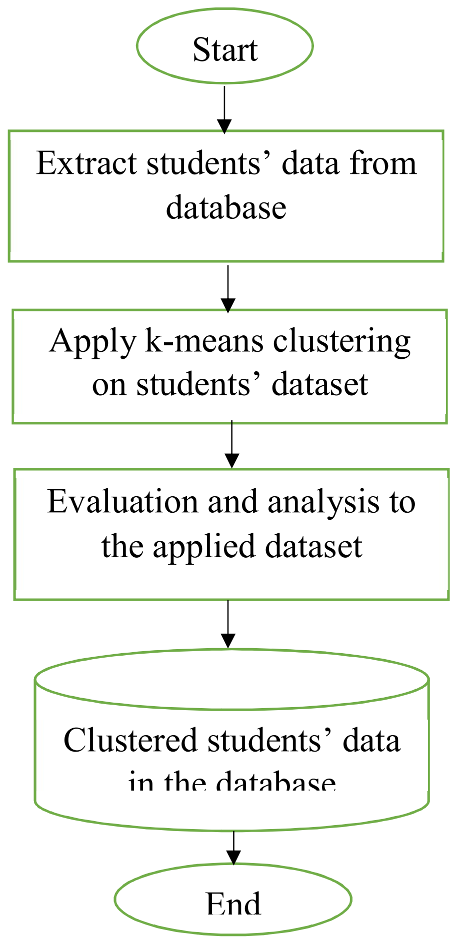

Figure 3.1 Flowchart of the proposed model

4. DATA PRESENTATION AND EXPERIMENTAL RESULTS

4.1 Data Preparations

The data set used in this study was obtained from the department of Computer Science, Federal University of Technology, Owerri. The initial size of the data is 50. In this step, data stored in different tables were joined in a single table after joining process errors were removed.

4.2 Data Selection and Transformation

Those fields which were required for data mining and a few derived variables were selected while some of the information for the variables were modified. All the predictor and response variables which were derived from the student dataset are given in Table 1 for reference.

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics Digital Humanities & Development Research
Vol 8, No 1, 2022 Series

Table 1: Student Related Variables

Variable	Description	Possible Values
PSM	Previous Semester Marks	[(Good > 60%) = 1], [(Average >36% & <60%) = 0.5)], [(Poor < 36%) = 0]
CTG	Class Test Grade	{Poor , Average, Good}
SEM	Seminar Performance	{Poor , Average, Good}
ASS	Assignment	{Yes = 1, No = 0}
GP	General Proficiency	{Yes = 1, No = 0}
*ATT	Attendance	{Poor , Average, Good}
LW	Lab Work	{Yes = 1, No}
ESM	End Semester Marks	(First > 60%), (Second >45 & <60%) (Third >36 & <45%) (Fail < 36%)

4.3 Experiments and Results

The results show the performance of the listed students by using data mining methodologies. Symbolic data analysis revealed that if students attends lectures regularly, frequently ask questions in class and do their assignments, they are more likely to earn good grades (probably overcoming the social-economic barrier) and gain promotion to the academic level without hitches. Mistakes that were associated together indicated to us that the very concept of formal proofs (i.e. the structure of each element of the proof, as opposed to the use of rules for instance) was a problem.

Surprisingly, results did not change much (a slight decrease in support and confidence levels in the following semester followed by a slight increase in performance). However, marks in the final exam continued increasing. This leads us to think that making mistakes, especially while using a training tool, is simply part of the learning process and was supported by the fact that the number of completed exercises per student increased.

a) The level of prediction seems to be much better when the prediction is based on exercises (number, length, variety of rules) rather than on mistakes made. This also supports the idea that mistakes are part of the learning process, especially in a practice tool where mistakes are not penalized.
b) Using data exploration and results from decision tree, one can infer that if students do successfully 2 to 3 exercises for the topic, then they seem to have grasped the concept of formal proof and are likely to perform well in the exam question related to that topic.

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html

Journal of Behavioral Informatics
Digital Humanities & Development Research
Vol 8, No 1, 2022 Series

This finding is coherent with correlations calculated between marks in the final exam and activity with the Logic Tutor and with the general, human perception of tutors in this course. Therefore, a sensible warning system could look as follows: to tutors, be proactive towards these students, distinguishing those who use out the pop-up menu for logic rules from the others.
Table 2 shows the students classification while table 4.2 shows the data set evaluation.

Table 2: Students classification

Student's Data	Details	Data	Value	Data2	Value2	Data3	Value3
Student's Basic Information	Serial Number	(1 - 50)					
	Student's ID						
	Date of Birth						
	Sex	M		F			
Socio-Economic Status	Status of Caregiver	Employed	1	Unemployed	0		
	Status of Student	Employed	1	Unemployed	0		
	Access to Basic Amenities	Yes	1	No	0		
Home or Community Factors	Home Residence	Formal	1	Informal	0		
	School Residence	On Campus	1	Off Campus	0		
	Attitude to Learning in Community	Above Aver	1	Below Average	0		
Teaching Environment Factors	Student's Perception of Lecture Venue	Good	1	Bad	0		
	Student's Perception of Tutor to Lecture Venue	Consistent	1	Inconsistent	0		
	Attitude of Student to attending Lectures	Consistent	1	Inconsistent	0		
General Attitude to Academic	Feeling about Success or Failure	Positive	1	Negative	0		
	Study Group Chat	Yes	1	No	0		
	Questions during Lecture	Yes	1	No	0		
	Number of Study Hours	Above 6 Hrs	1	Below 6 Hrs	0		
	Struggle to Concentrate	yes	1	No	0		
Academic Performance	Past Semester Marks	PSM>60%	1	(45%<PSM<60%	0.5	PSM<45%	0
	Class Test Grade	PSM>60%	1	(45%<PSM<60%	0.5	PSM<45%	0
	Seminar	Good	1	Average	0.5	Poor	0
	Assignment	Yes		Average	0.5	Poor	0
	General Proficiency	Good	1	Average	0.5	Poor	0
	Attendance	Regular	1	Irregular	0.5	Poor Attendance	0
	Lab Work	Good	1	Average	0.5	Poor	0
Prediction	End of Semester Marks	Excellent		Average		poor	

Creative Research Publishers
https://www.crossref.org/06members/50go-live.html
Journal of Behavioral Informatics, Digital Humanities & Development Research
Vol 8, No 1, 2022 Series

Table 3: Data Set Evaluation

S. No.	PSM	CTG	SEM	ASS	GP	ATT	LW	ESM
1.	First	Good	Good	Yes	Yes	Good	Yes	First
2.	First	Good	Average	Yes	No	Good	Yes	First
3.	First	Good	Average	No	No	Average	No	First
4.	First	Average	Good	No	No	Good	Yes	First
5.	First	Average	Average	No	Yes	Good	Yes	First
6.	First	Poor	Average	No	No	Average	Yes	First
7.	First	Poor	Average	No	No	Poor	Yes	Second
8.	First	Average	Poor	Yes	Yes	Average	No	First
9.	First	Poor	Poor	No	No	Poor	No	Third
10.	First	Average	Average	Yes	Yes	Good	No	First
11.	Second	Good	Good	Yes	Yes	Good	Yes	First
12.	Second	Good	Average	Yes	Yes	Good	Yes	First
13.	Second	Good	Average	Yes	No	Good	No	First
14.	Second	Average	Good	Yes	Yes	Good	No	First
15.	Second	Good	Average	Yes	Yes	Average	Yes	First
16.	Second	Good	Average	Yes	Yes	Poor	Yes	Second
17.	Second	Average	Average	Yes	Yes	Good	Yes	Second
18.	Second	Average	Average	Yes	Yes	Poor	Yes	Second
19.	Second	Poor	Average	No	Yes	Good	Yes	Second
20.	Second	Average	Poor	Yes	No	Average	Yes	Second
21.	Second	Poor	Average	No	Yes	Poor	No	Third
22.	Second	Poor	Poor	Yes	Yes	Average	Yes	Third
23.	Second	Poor	Poor	No	No	Average	Yes	Third
24.	Second	Poor	Poor	Yes	Yes	Good	Yes	Second
25.	Second	Poor	Poor	Yes	Yes	Poor	Yes	Third
26.	Second	Poor	Poor	No	No	Poor	Yes	Fail
27.	Third	Good	Good	Yes	Yes	Good	Yes	First
28.	Third	Average	Good	Yes	Yes	Good	Yes	Second
29.	Third	Good	Average	Yes	Yes	Good	Yes	Second
30.	Third	Good	Good	Yes	Yes	Average	Yes	Second
31.	Third	Good	Good	No	No	Good	Yes	Second
32.	Third	Average	Average	Yes	Yes	Good	Yes	Second
33.	Third	Average	Average	No	Yes	Average	Yes	Third
34.	Third	Average	Good	No	No	Good	Yes	Third
35.	Third	Good	Average	No	Yes	Average	Yes	Third
36.	Third	Average	Poor	Yes	No	Average	Yes	Third
37.	Third	Poor	Average	No	Yes	Average	Yes	Third
38.	Third	Poor	Average	No	No	Poor	Yes	Fail
39.	Third	Average	Average	No	No	Poor	Yes	Third
40.	Third	Poor	Poor	No	No	Good	No	Third
41.	Third	Poor	Poor	Yes	No	Poor	Yes	Fail
42.	Third	Poor	Poor	No	No	Poor	No	Fail
43.	Fail	Good	Good	No	Yes	Good	Yes	Second
44.	Fail	Good	Good	Yes	Yes	Average	Yes	Second
45.	Fail	Average	Good	No	Yes	Average	Yes	Third
46.	Fail	Poor	Poor	No	Yes	Average	No	Fail
47.	Fail	Good	Poor	No	No	Poor	Yes	Fail
48.	Fail	Poor	Poor	No	No	Poor	Yes	Fail
49.	Fail	Average	Average	Yes	Yes	Good	Yes	Second
50.	Fail	Poor	Good	Yes	No	Poor	No	Fail

4. CONCLUSION

K-mean algorithm for clustering dataset plays a vital role in the concept of data mining. The algorithm was used to cluster students' datasets and identify recognized pattern that group students' performances to see how far they can succeed in an academic session. The application of K-mean clustering algorithm on these datasets proves useful in avoiding ambiguities when grouping the datasets for use in the education sector.

In this work, the clustering task is used on student database to predict the students' performance through various divisions on the basis of selected appropriate factors. Information like attendance, class test, seminar and assignment marks were collected from the students' previous database to predict the performance at the end of the semester. This study will be of help to the students and the teachers to improve the divisions encountered most times between teachers and their students. This study will also work to identify those students who needed special attention to reduce the number of weak students, and taking appropriate action for the next semester examination.

5. RECOMMENDATION FOR FUTURE WORK

One limitation of the algorithm used is that the value of k; the number of clusters, is still required to give as an input, regardless of the distribution of the data points. Recommendation for future work includes evolving some statistical methods to compute the value of k, depending on the data distribution. Methods for refining the computations of initial centroids are worth investigating. Furthermore, it is also recommended that an interface be designed for easy input and output of students' record and also a database that can hold the information of a large body of students be designed.

REFERENCES

1. Brojo Kishore Mishra, D. H. (2016) Business Intelligence using Data Mining Techniques and Business Analytics. *5th International Conference on System Modeling & Advancement in Research Trends* 2(3), 86.
2. Harun Bayer, M. A. (2017) Big Data mining and Business Intelligence Trends. *Journal of Asian Business Strategy* 7(1), 23-33.
3. Han J and Kamber M (2000). Data Mining: Concepts and Techniques," Morgan Kaufmann. 7(2) 45-46.
4. Pandey U.K and Pal S (2011). Data Mining: A Prediction of Performer or Underperformer using Classification. *International Journal of Computer Science and Information Technology*, 2(2), 686-690.
5. Hijazi S.T, and Naqvi R.S.M.M (2006) Factors Affecting Student's Performance: A Case of Private Colleges, *Bangladesh e-Journal of Sociology*, 3(1), 6-7.
6. Khan Z.N (2005) Scholastic achievement of higher secondary students in science stream, *Journal of Social Sciences*, 1(2), 84-87.
7. Galit (2007) Examining Online Learning Processes Based on Log Files Analysis: a case study. Research, Reflection and Innovations in Integrating ICT in Education 1-5.
8. Al-Radaideh Q.A, Al-Shawakfa E.W, and Al-Najjar M.I (2006) Mining student data using decision trees, *International Arab Conference on Information Technology* 1-5.