International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 4 Issue 1, January 2017

ISSN: 2394-4404

Educational Data Mining And Its Applications

Mistura M. Usman

Adamu Y. Atumoshi

Department of Computer Science,
University of Abuja, Gwagwalada, Abuja, Nigeria

Abstract: ***The increasing interest in educational data mining makes it a new growing research area. Educational data mining also referred to as "EDM" is defined as the area of scientific inquiry centered around the development of techniques for making discoveries within the sets of data found in learning environment, and using those techniques to understand the students and the learning environment. It is a very powerful tool to reveal hidden patterns and precious knowledge, which otherwise may be difficult to establish and comprehend using traditional statistical methods. This paper introduces educational data mining, its advantages and limitation. discusses educational data mining techniques by conducting step by step processes also identified the potential areas in which data mining techniques can be applied in the field of Higher education and the data mining technique that is suited for such application.***

Keyword: Educational Data, Educational data Mining, Prediction, Clustering, Association rule.

I. INTRODUCTION

The increasing use of technology in educational system has led to a tremendous change in the way educational systems operate. The increased use of electronic based learning systems has also amplified the amount of data available and more readily accessible for decisions making. The improvements in the data mining algorithms also make analysis of this volume of data easier. In recent times, there has been increased interest and research in the field of Educational Data Mining [1].

With the increasing research interests in the use of data mining in educational system and with the emerging field of Educational Data mining, which is concerned with developing methods that discover knowledge from data originating from educational Institutes [2].

These data can be collected from different educational system where the data reside in their databases. The data can be personal records or academic records that can be used to understand students' learning behavior, to assist instructors, in improving teaching, can also be used to evaluate and improve e-learning systems, to improve curriculums and the for overall decision making.[2][3].

A. EDUCATIONAL DATA

Reference [4][5], Educational data are data generated from different sources such as diverse and distributed, structured and unstructured data. They could either be from offline or online sources as shown in fig.1. Offline Data are generated from traditional classroom activities, Interactive teaching / learning environments, learner / educators information, students attendance, Emotional data, Course information, data collected from the academic section of an institution. While the Online Data, are generated from the distance learning education system, web based learning system, geographically separated stake holder of the education, and computer supported collaborative learning used in social networking sites and online forum. Like : Web logs, E-mail, Text data, Transcripted Telephonic Conversations, Medical records, Spreadsheets, Corporate contracts, publication databases, Legal Information[4][5]

ISSN: 2394-4404

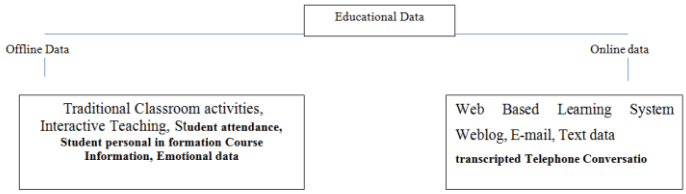

Figure 1: Educational data

B. EDUCATIONAL DATA MINING

Reference [6] defined EDM as "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in"

Reference [7] defined Educational data mining (that is "EDM") "as the area of scientific inquiry centered around the development of methods for making discoveries within the unique kinds of data that come from educational system, and using those methods to better understand students and the system where they learn" Educational data mining techniques often differ from techniques of the broader data mining literature, in explicitly exploiting the multiple levels of meaningful hierarchy in educational data. Methods from the psychometrics literature are often integrated with methods from the machine learning and data mining literatures to achieve this goal [7].

Education Data Mining (EDM) is the application of data mining techniques relating for learning analytics and quantitative observation method in order to understand how student respond to educational system and how their responses impact their learning. Its objective is to analyze educational data in order to resolve educational research problem. In recent years there is rapid growth in education sector which leads to growing of education data, so educational data mining become important to understand student learning behavior during learning process and to understand their challenges [8].

From the fig2. above in EDM system. The educationists worked upon the educational system to ensure the performance of students. as shown that :

- ✓ The academician will design the educational system, build the system and most importantly maintains that educational system. These educational the systems include traditional classrooms and some E- learning system, intelligent and adaptive web based educational system etc.
- ✓ The data set can be extracted from learner as are directly connected with educational system.
- ✓ these data is given as input to data mining processes from which results are given as recommendations to learner and to discover new knowledge to the educators by using various data mining techniques like clustering, classification, association rule etc [8].

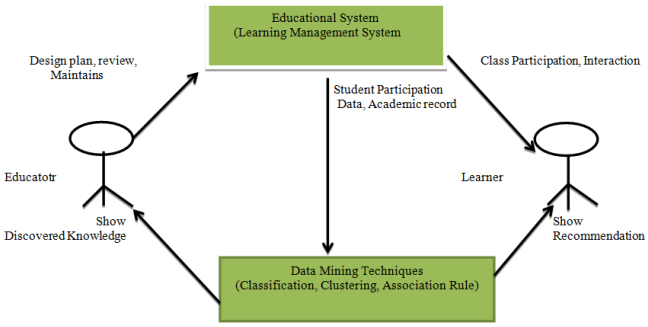

Figure 2: Educational data mining system

C. GOALS OF EDUCATIONAL DATA MINING

The major goal of EDM is to improve the overall educational system. All these goals depend on the view point of the end users, which includes: student, educator, administrator and researcher and to help solve their problems [5].

- ✓ Student Modeling: building the student model which incorporates detailed information such as student's learning progress, student's characteristics such as knowledge, skills, motivation, satisfaction, meta-cognition, attitudes, experiences and or certain types of problems that negatively impact their learning outcomes. The goal here is to create or improve student model
- ✓ Predicting students' future learning behavior: through the predictive models students' performance and learning outcomes can be predicted based on data from course activities.
- ✓ Making Recommendations: The goal is to recommend most appropriate content or tasks at the current time to student.
- ✓ Studying the effects of educational support and suggesting further educational support system
- ✓ It can be achieved with the help of learning systems,
- ✓ Allowing educators to understand their student's learning processes and updating their teaching methods, also, to improve teaching/ learner performance, to understand social, cognitive and behavioral aspects
- ✓ Advancing the general scientific knowledge on learning and learners- can be achieved by building computational models which incorporates the student models, technology and software are also used in the area of EDM research
- ✓ Discovering or improving domain models: The goal here is to determine how to improve courses or contents, activities, links, etc.), using information (in particular) about student usage and learning. Learners are engaged in educational content in determining optimal instructional sequences in order to support the learning style of students
- ✓ Administrators to understand the best way to extend the institutional resources both human and material for the overall development of educational system [4].

D. LIMITATION OF DATA MINING

The research trend of Educational Data Mining shows that maximum research focuses only on academic objectives. The other issues are: [5].

- ✓ Incremental nature of educational data: The exponential increase of data makes the maintenance of the data in data warehouse very difficult. Also in the monitoring of the operational data sources, infer the student interest, intentions and its impact in a particular institution is the main issue. There is also the issue of the alignment and interpretation of the incremental data. It should focus on appropriating time, context and its sequence. Another issue of incremental is the optimal utilization of human computing and resources [5].
- ✓ Occurrence of Uncertainty: The presence of uncertain errors, no model can accurately give a prediction of hundred percent results of student model or the general academic planning [5].
- ✓ Problem of very high dimensional data : dealing with very large databases and very high dimensionality need to be resolved; over- fitting existing data, missing and noisy data, and. Furthermore, for large-scale, real-world tasks, high performing algorithms such as neural networks and genetic algorithms must cope with long computation times and difficulties in making interpretations [9].
- ✓ Techniques associated with probabilistic learning require to be improved. For educational data mining and analytics in classrooms, schools, districts, and other institutions to be successful, the techniques associated with probabilistic learning needs to be improved, this will enable student to pose questions that matter to teachers and other users and to frame findings in a thoughtful, informative way that highlights and recommends clear actions. In reports about the newest technologies for adaptation, personalization, and recommendation, the role of human judgment is sometimes underemphasized with the exception of visual data analytics [9].
- ✓ Research Expertise Relation between Student/Teacher. In most of the higher Educational institutions supervisors are assigned based on the availability and area of expertise in the different department, With this, it is not possible to assign all the students /supervisor with similar area interest area, hence this may lead to having result of the project that is not applicable to real scenarios. There is a need to find the relation between areas of interest, students' interest, and applicability of the project/research and mining cross interest. introducing Association Mining will be beneficial in optimizing this issue [4].

E. PHASES IN EDUCATIONAL DATA MINING

Phases of Educational Data Mining Educational Data Mining are concerned with translation of raw data collected from educational systems to discovering a new hidden information. Generally EDM consist of following phases, as shown in fig.3 below

- ✓ Data collection phase : The data to be mined is collected from different educational system resources i.e. from course management, E-learning environment, web based data (i.e. YouTube, twitter)which is relevant to students activities during learning process like their academic grades, students posts on social networking sites etc
- ✓ Data Mining Phase: The data mining phase consist of the following process phases
  - Discover Relationship: The first phase of educational data mining is to find the relationships between the data of educational environment using data mining techniques i.e. classification, clustering, regression etc. with the objective of finding consistent relationships between data variables
  - Validating Relationships in the second phase of educational data mining, validation of discovered relationships between data are done so that uncertainty can be avoided and in order to avoid over fitting
  - Making Predictions: The third phase is to make predictions for future on the basis of validated relationships in learning environment. Relationships which are valid are used to make predictions about further events in the learning environment
  - Decision Making: The fourth phase is supporting decision making process with the help of predictions. Predictions made in previous phase are used to support decision-making processes and in making policy decisions. During phases 3 and 4, data is often visualized or in some other way to make distillation of human judgment [8]

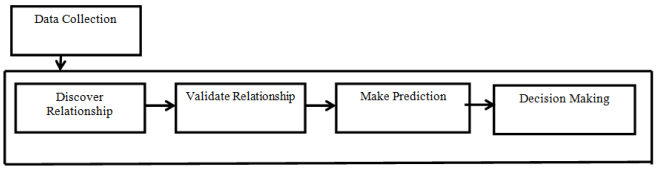

Figure 3: Educational data mining process phase

II. MATERIALS AND METHODS

A. EDUCATIONAL DATA MINING TECHNIQUES

According to [5], Educational Data Mining does not only apply data mining techniques: Classification, clustering, and association analysis, but it also apply other methods and techniques drawn from the different areas related to EDM (statistics, machine learning, text mining, web log analysis, etc.). There are so many methods of educational data mining, to classify them, educational data mining researchers uses these five categories of technical methods: [7]

- ✓ Prediction: in prediction, a model is developed to infer a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables). Broadly, there are three types of prediction: classification ((IF-THEN) rules, decision trees), regression, and density estimation. In classification, the predicted variable is a binary or categorical variable. Some popular classification methods include decision trees, logistic regression (for binary predictions), and support vector machines. In regression, the predicted variable is a continuous variable. Some popular regression

ISSN: 2394-4404

methods within educational data mining include linear regression, neural networks, and support vector machine regression. In density estimation, the predicted variable is a probability density function. Density estimators can be based on a variety of kernel functions, including Gaussian functions. For each type of prediction, the input variables can be either categorical or continuous; different prediction methods are more effective, depending on the type of input variables used [7].

- ✓ Clustering: In clustering technique, the data set is divided in various groups, known as clusters. As per clustering phenomenon, the data point of one cluster and should be more similar to other data points of same cluster and more dissimilar to data points of another cluster. There are two ways of initiation of clustering algorithm: Firstly, start the clustering algorithm with no prior assumption and second is to start clustering algorithm with a prior postulate [8][10].
- ✓ Relationship mining: It is used for discovering relationships between variables in a dataset and encoding them as rules for later use. There are different types of relationship in mining techniques such as association rule mining (any relationships between variables), sequential pattern mining (temporal associations between variables), correlation mining (linear correlations between variables), and causal data mining (causal relationships between variables). In EDM, relationship mining is used to identify relationships between the student's on-line activities and the final marks and to model learner's problem solving activity sequences [8][10].
- ✓ Discovery with Models: the goal of discovery with models is to use a validated model of a phenomenon (using prediction, clustering, or knowledge engineering) as a component in further analysis such as prediction or relationship mining. It is used for example to identify the relationships between the student's behavior and characteristics [8][10].
- ✓ Distillation of Data for Human Judgment: In this case, human beings can make inferences about data, when it is presented appropriately, that are beyond the immediate scope of fully automated data mining methods. The methods used in this area of educational data mining are information visualization methods.

Data is distilled for human judgment for two reasons: that is, identification and classification. When data is distilled for identification, it is been displayed in

Ways that enables human to easily identify known patterns that are not too difficult to formally express. Example of educational data mining visualization is the learning curve.

Data may also be distilled for human labeling, to support the later development of a prediction model. Here, sub-sections of a data set are displayed in visual or text format and labeled. These labels are then serves as the basis for the development of a predictor [7].

B. APPLICATION AREAS

The potential area of Educational data mining application includes:

- ✓ PREDICTION OF STUDENTS ENROLMENT INTO PROGRAMS

Educational data mining can be applied to find an accurate estimate of how many male or female will enrol in a particular program by using the Prediction techniques this will in-turn lead to efficiency in allocation of resources.

- ✓ PREDICTING STUDENT PERFORMANCE

Many researchers used different data mining techniques to predict student performance. In recent times, learning is taking on an important role in the development of our civilization. Learning is an individual behavior as well as a social phenomenon. With the help of data mining techniques a result evaluation system can be developed which can help teachers and students to know the weak points of the traditional classroom teaching model. Also it will help them to face the rapidly developing real-life environment and adapt the current teaching realities [11]. It is too tedious to investigate and successfully propose models for evaluating learning efforts with the combination of theory and practice. the goals of University is clearly to promote teaching and learning through effective teaching, research and scholarship in service to the university Student learning is considered in some goals and outcomes related to the development of overall student knowledge, skill, and dispositions. Collections of randomly selected student work are examined and assessed by small groups of faculty teaching courses within some general education categories [12].

- ✓ ORGANIZATION OF SYLLABUS

It is important for educational institutes to maintain a high quality educational programme which will improve the overall learning process and will help the institute to optimize the use of resources. A typical student at the university is expected to complete a number of courses before to graduation and in order to facilitate students' learning capacity optimally, exploration of subjects and their relationships can directly assist in better organization of syllabi and provide insights to existing curricula of educational programmes. One of the applications of data mining is to identify related subjects in syllabi of educational programmes in a large educational system[13]

- ✓ ERRONEOUS/ABNORMAL/VALUES IDENTIFICATION

The data stored in a database may reflect outliers-|noise, or incomplete data objects, exceptional cases, which may confuse the entire analysis process, leading to over fitting of the data to the knowledge of the model constructed. As a result, the discovered pattern may not be accurate [14]. One area of applications of the Outlier Analysis is to detect an abnormal values in the student's result sheet

✓ DETECTING CHEATING IN ONLINE EXAMINATION

Online assessments can be used to evaluate students' knowledge, they are used around the world in elementary schools to institutions of higher learning and other recognized training centers like the Cisco Academy [15].

In recent times examinations are conducted online remotely through the Internet and if a fraud occurred , it can be detected by tracing the identity of the who was there at that particular time, cheating is not only done by students but the current scandals in journalism and business shows that it is now a common practice. Data mining techniques are used to propose models which can help organizations to detect and to prevent cheats in online assessments. The models proposed use data comprising of common student's personalities, stress situations generated by online assessments, and common traits used by students to cheat in order to obtain higher grades in their examinations [11]

III. RESULTS AND DISCUSSION

In the University of Abuja, where you have enrolment in to a program as show in Figure4, using prediction techniques we can predict the students' future enrolment.

Year	Number of male	Number of Female
2010	60	35
2011	50	30
2012	65	45
2013	55	30
2014	70	50
2015	80	?

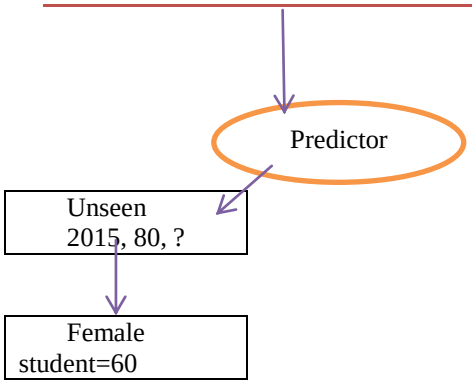

Figure 4: predicting Female Enrolment

In university of Abuja, student participation data in activities are used in assessing performance and the quality of student

All the predictor and response variables which were derived from the database are given in the table 1.

Variables	Description	Values
Midterm	Midterm test	Excellent>=80 Very Good >=65 &< 80 Good >=50 & < 65 Acceptable >= 45 & < 50 Fail < 45
LAB	Lab test	(Poor, Average, Good)
ASS	Attempt Assignment	(Yes,No)
CP	Class Participation	(Yes,No)
SEM	Seminar performance	(Yes,No)
ATT	Student Attendance	(Poor, Average, Good)
FGS	Final Grade Scored	Excellent>=80 Very Good >=65 &< 80 Good >=50 & < 65 Acceptable >= 45 & < 50 Fail < 45
CD	Class of Degree	(Good, Acceptable)

Table 1: Student Variable Table

Student Id.	Midterm	LAB	ASS	CP	SEM	ATT	FGS	CD
001	Excellent	Good	Yes	Yes	Good	Good	Excellent	Good
002	Excellent	Good	Yes	No	Average	Good	Excellent	Acceptable
003	Good	Good	No	Yes	Average	Average	Very Good	Acceptable
004	Excellent	Good	No	Yes	Good	Good	Excellent	Good
005	Excellent	Average	Yes	Yes	Good	Good	Excellent	Good
006	Very Good	Average	No	No	Good	Average	Good	Acceptable

Table 2: Student Participation Table

Set of Decision Rule Generated from table

IF Midterm='Excellent' AND LAB= 'Good' AND CP='Yes' AND SEM='Good' AND ATT= 'Good' AND ASS='Yes' AND FGS='Excellent' THEN CD='Good'

IF Midterm='Excellent' AND LAB= 'Good' AND CP= 'NO' AND SEM='Average' AND ATT= 'Good' AND ASS='Yes' AND FG ='Excellent' THEN CD='Acceptable'

IF Midterm='Very Good' AND LAB= 'Average' AND CP= 'NO' AND SEM='Good' AND ATT= 'Average' AND ASS='NO' AND FG ='Good' THEN CD='Acceptable'

A study was conducted to find the strongly related subjects in a course offered by the student of the university, to find this: association rule mining was used to identify possible related two subject combinations in the syllabi which also reduce the search space, then Pearson Correlation Coefficient was applied to determine the strength of the relationships of the identified subject combination.

Student Identity	**Course I**	**Course II**	**Course III**
001	Database design and management	programming I	programming II
002	Database design and management	programming I	programming II
003	Database design and management	programming I	programming II
004	Database design and management	programming I	Microprocessor
005	Database design and	programming I	Computer Networks

	management		

Table 3: Course Combination Table

Association Rules from Table 3
the form:

(X, course1) ⇒ (X, course 2) (1)

(X, course 1) ^(X, course 2) ⇒ (X, subject3) (2)

(X," Database design and management")⇒ (X," programming I")
[support=2% and confidence=60%] (3)

(X," Database design and management ")^ (X," programming I ")⇒ (X," programming II")
[support=1% and confidence=50%] (4)

Where support factor of the association rule shows that 1% of the students have taken both the subjects "Database design and management" and "programming I" and the confidence value shows that there is likely 50% chance that the students who took "Database design and management" will also take "programming I"

This way we can find the strongly related subjects and can optimize the syllabi of an educational programme

The Outlier Analysis is used to detect an abnormal values in the student's result sheet which may be due factors such human error, software malfunction, or xtraordinary performance from the student in that course as shown in Table 4.

Student Id.	Score1	Score2	Score3	Score4	Score5
001	35	40	35	45	30
002	65	64	60	71	75
003	90	85	79	80	75
004	50	48	55	50	57
005	35	30	45	40	**98**

Table 4: Student Scores

In the table shown above the result of the student in Score5 with student Id. 005 will be identified as an exceptional case and can be further analyzed for the cause.

IV. CONCLUSION

The application of data mining techniques in institute of higher learning brings a lot of advantages in the area of decision making, the paper, discussed the various data mining techniques which can support education system, we have shown that if these techniques are applied they will help in prediction of student performance and enrolment of student into programmes, continuous retainment of academic programme, and organization of syllabus in the in the institution for optimal allocation of resources,

REFERENCES

[1] C. Romero, S. Ventura "Educational data Mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp.135-146, 2007

[2] C. Romero, S. Ventura, E. Garcia, "Datamining in course management systems: Moodle case study and tutorial", Computers & Education, Vol. 51, No. 1, pp. 368-384, 2008

[3] C. Romero, and S. Ventura, "Educational data mining: a review of the state of the art," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 40, no. 6, pp. 601-618, 2010.

[4] R.Jindal, M.D Borah, A Survey on Educational Data Mining and Research trends, International Journal of Database Management System (IJDMS), 5(3), 2013, 53–73.

[5] Kashyap1 G. and Chauhan E. " Review on Educational Data Mining Techniques " International Journal of Advanced Technology in Engineering and Science Vol. No.3 Issue 11, November 2015 pp. 308-316

[6] Baker, R., & Yacef, K. (2009). The State of Educational Data mining in 2009: A Review Future Visions. Journal of Educational Data Mining, 1 (1)

[7] Baker, R. S. J. d. 2011. "Data Mining for Education." In International Encyclopedia of Education, 3rd ed., edited by B. McGaw, P. Peterson, and E. Baker. Oxford, UK: Elsevier.

[8] Upadhyay N. and Katiyar V. "A Survey on the Classification Techniques in Educational Data Mining", International Journal of Database Management System (IJDMS), 3(11), 2014, 725–728

[9] Ritu Gautam, Deepika Pahuja A Review on Educational Data Mining International Journal of Science and Research Volume 3 Issue 11, November 2014

[10] Baradwaj, B. Kumar, Mining Educational Data to Analyze Students Performance. International Journal of Advanced Computer Science and Applications (IJACSA), 2(6), 2011, 63-69.

[11] Varun Kumar and Anupama Chadha An Empirical Study of the Applications of Data Mining Techniques in Higher Education International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011

[12] Sun Hongjie, "Research on Student Learning Result System based on Data Mining", IJCSNS International Journal of Computer Science and Network Security, Vol.10, No. 4, April 2010

[13] Tissera, Athauda, and Fernando 2006). "Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining", IEEE International Conference on Information Acquisition, 2006

[14] Han Jiawei, Micheline Kamber, Data Mining: Concepts and Technique.Morgan Kaufmann Publishers,2000

[15] Academy Connection – Training Resources In html, http://www.cisco.com/web/learning/netacad/index