Reflection Paper

iMedPub Journals
http://www.imedpub.com/

American Journal of Computer Science and Information Technology

2019
Vol.7 No.2:37

Data Collection Experience on Educational Data Mining in Nigeria

Ebiemi Allen Ekubo*

Department of Computer Science and Information Systems, North-West University, Mafikeng Campus, South Africa

***Corresponding author:** Ebiemi Allen Ekubo, Department of Computer Science and Information Systems, North-West University, Mafikeng Campus, South Africa, E-mail: ebiemi4allen@gmail.com

Received date: June 20, 2019; **Accepted date:** June 26, 2019; **Published date:** July 3, 2019

Citation: Ekubo EA (2019) Data Collection Experience on Educational Data Mining in Nigeria. Am J Compt Sci Inform Technol Vol.7 No.2: 37

Abstract

Educational data mining and learning analytics encourages developing methods for discovering student learning patterns and behaviors by investigating the distinct set of data available in learning environments. Researchers in this field of domain have developed useful models by exploring data from different learning environments. Most of the data used by these researchers come from computer based learning environment where datasets can be easily fetched and analyzed. For the goal of educational data mining to be fully realized, researchers must engage in investigating data from every learning environment, either in computer based learning environment or traditional learning environment; or institutions with information management systems or not. With is this view in mind, this research aim to expose the experience during data collection at two public universities in Bayelsa state, Nigeria. The research describes the data collection process and highlights the challenges faced. The research concludes by encouraging researchers to share their data collection experiences to help future researchers to be adequately prepared and encourage educational data mining researchers to support the goals of the educational data mining community by studying students from every learning environment.

Keywords: Data collection; Educational data mining; Data cleaning

Introduction

Data mining in education is concerned with the exploration of the unique set of data available within the education background and making use of the acquired knowledge to understand and improve students' learning outcome and their learning environment [1]. From the reviews of Baker and Yacef and Romero and Ventura some areas where EDM is successfully applied are improvement of student models, analysis of domain structures, prediction of students' performance, identifying student learning behaviours and personalizing student recommendation system [2,3]. The following sources for analysing educational data highlighted in [4] include Learning Management Systems, Information systems used in managing student information, Intelligent Tutoring systems for training specific skills with data available at online repository, social media, questionnaires or forums online. Romero and Ventura broadly describes traditional education and computer-based educational systems as the types of educational environments; noting that while traditional education focuses on apply traditional methods of teaching and student data collection (although some form of computer systems can be incorporated in this method); computer based education solely makes use of computers for instructing and managing student information [5]. The computer-based educational systems store student information in file logs; with access to these files, analysis done can improve aspects of the system and provide students with the help they require [6,7]. For traditional education, access to student information requires gathering information from different sources, even in cases where institutions make use of information systems, educators have to observe and note down student behaviours and learning outcome, although this can be tedious, knowledge from this learning environment is required to improve learning within this area. Even with the developments in educational technology, many higher institutions in developing countries are yet to incorporate these technologies into their education system [8,9] . In Nigeria, many public universities make use of computers only as a means of storing student details and results, with little information on student learning process or behaviour. This research looks at the available recorded data at two of such universities in Nigeria with the aim of revealing the data collection experience. These institutions owned by the state and federal governments are located in Bayelsa state. The paper describes the entire data collection process, highlights the challenges witnessed, offers discussions on the dataset and the process of gathering the data, and provides recommendations for future studies.

Literature Review

Data collection or gathering in educational data mining has majorly focused on data available at online data repositories. From the survey of Romero et al. online platforms are data banks for gathering educational data [10]. Platforms such as Massive Open Online Course (MOOC), Intelligent Tutoring System (ITS) and Learning Management Systems produce gigantic data on regular basis that is easily accessible for mining [11-14].

Some researchers in their work describe the learning environment of the dataset used in their research. For example, in the work of Mueen et al. they collected data from a learning management system for undergraduate students who took two programming courses over a period of two semesters to predict and analyse students' academic performance from their academic records and forum participation [15,16]. Ahmed and Elaraby obtained dataset from a student's database in one department to predict student performance in the department [17,18]. Costa et al. analyzed the effectiveness of educational data mining techniques in identifying academic failure in introductory programming by analysing dataset extracted from two sources, an online distance education system and an on-campus programming course with information about student information details and course activities [14]. Hernández-García et al. also described the data source environment as Moodle database and stated that they used an extraction, transformation and loading process to provide the pointers they needed in the research to predict the group assessment for teamwork [19].

Other researchers described the dataset but failed to discuss the learning environment where they collected the data. In their work, Guarín et al. mentioned the institution where they collected the data and gave detailed description of the dataset with no detail of the environment where they got the dataset [16]. Oyerinde and Chia in their work of predicting students' academic performance using multiple linear regression barely described the data source environment by stating the name of the data source [12]. In the work by Mythili and Shanavas they stated the source of their data as database, with no information about the institution or learning environment [18]. The work of Yassien et al. described the dataset and only stated the institution of the data source with no mention of the learning environment [20].

The review above shows that while some researchers explain their data source environment; others focus on describing the set of data used but fail to describe the environment of the data source. Although, the dataset is more important in discovering knowledge and patterns, it is also important to understand and explain the environment where these dataset originates from to promote the growth of educational data mining in every learning environment.

Method

- The research follows the method depicted with the aid of the diagram labelled in Figure 1. The diagram shows the entire process followed and below is a brief description of the process:
- Distribute letters: This step handles the distribution of letters to the universities to seek their permission for the gathering of data from the institutions.
- Meet with stakeholders: Meeting with the authority of the universities to discuss the pattern of data collection, this meeting allowed the researcher and the stakeholders to agree on terms regarding the data offered and privacy for the students.
- Collect secondary data from available sources: This is the data collection stage, where the researcher engaged in gathering secondary information available in both universities from different locations and sources.
- Distribute questionnaire to low performing students: This involves distributing questionnaires to low performing students to gain more information about them that is not available in the secondary information collected at the university.
- Collect responses from students: These are the collected responses from the low performing students, aiding in the provision of relevant features for the research.
- Collate data into one format: This combines the secondary data collected in different formats and sources, and the primary data collection from the low performing students. This step involves collating all the data into manageable files using the Microsoft Excel spreadsheet package.
- Clean data: After collating the data into one format, this stage involves ridding the data of incomplete records and handling errors found in the data. Cleaning of data is required in data mining to ensure the data has quality.
- Data set for mining: Data set for mining is not a step but rather a product or output of the cleaned data from all the data gathered. This is the dataset ready for knowledge mining.

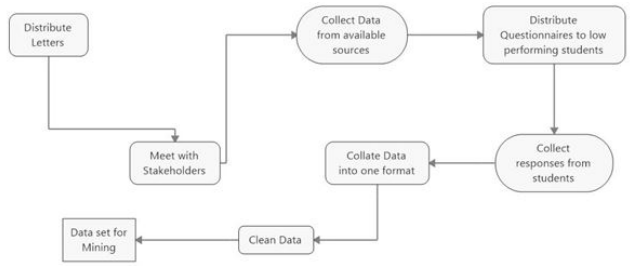

Figure 1: Data collection process.

Data collection preparation

- To prepare for the data collection, the researcher engaged in the following:
- Gaining ethical approval for the research: For the research ethical approval stage, after presentation and acceptance of the research proposal, the school ethics committee granted approval to carry out the research work.
- Obtaining letters for distribution to the universities for data collection: The researcher obtained letters for distribution to the universities from dean of the faculty.
- Distributing and obtaining approval from the universities to commence data collection: These letters were distributed and an ample amount of time given while waiting for feedback from the universities.

With feedback and approval from the universities, the researcher embarked on the data collection journey. The result from this experience explained in detail follows in the next section.

Results

The research carried in the two public universities shows two different pattern of data storage and access. The state owned university has a faculty officer and an examination officer; the faculty officers handle the storage of student data, which are available in spreadsheet format while the examination officers handle the storage of student results available in printed hard paper copy or pdf format. The federal owned university has an examination office that handles the storage of student records and the data are available in spreadsheet format, the student record available for this institution are the records of final year students alone, students of different level only had result information.

The state owned university ensured that students provide and store all their details during registration in their first year and continually update this information as they progress in their study. According to the information relayed to the researcher by the head examination officer, the federal owned university deems it fit to collect these details for storage at the final level of students to enable them process students graduating for that session for the National Youth Service Corps (NYSC) programme.

Document gathering

The data gathering process were different for both universities. The time spent on gathering the data for both universities took about seven weeks from the letter distribution to the data collection stage.

The federal owned university with the examination office received the memo sent by the registrar of the institution to assist the researcher by sharing all the available information ensuring the identity of the students are kept private. With the authorization received from the registrar, the head exam officer directed the junior exam officer to transfer the available data to the researcher. The office makes use of excel spreadsheet to store all the student details and result information. As the researcher waited, the junior officer removed all information that could easily identify students and gave the data to the researcher through a USB flash drive. To ensure the data is in the device, the researcher checked through the researcher's own laptop to confirm. The researcher noticed the small amount of data and enquired about it; the head exam officer made the researcher understand that the data available at the examination office is for final year graduating students alone.

For the state owned university, the researcher met with the deputy vice chancellor academics (DVC Acad.) of the university who is in charge of handling all information related to students' academics. The DVC Acad. assisted the researcher by sending out memos to all faculty deans to assist the researcher in the data gathering. The deans assisted by sending out memos to their heads of departments, faculty officers and faculty examination officers. With all deans and faculty sections notified, the intention of the researcher was to collect data from one faculty and then move to the next faculty. However, the researcher experienced delays in response from faculty officers and examination officers and decided to combine two or three faculties depending on the rate of response. Meeting with each faculty officer was mostly productive on the same day with few exceptions where the faculty officers postponed the data collection for a later date. The faculty officers stored the student details in spreadsheet files and gave the researcher the data on USB flash drive. The student result data collection involved meeting with examination officers within departments; for some departments where the examination officers were unavailable, the heads of departments provided the data, which were either in pdf files or paper hard copy files. The researcher transferred each file collected through USB flash drive immediately into the researcher's laptop and stored all paper hard copy files in a file jacket.

At the end of the data collection, the researcher gathered all the files received from both universities and stored them in two different folders for collating and cleaning.

At the end of the data collection, the researcher gathered all the files received from both universities and stored them in two different folders for collating and cleaning.

Questionnaire gathering

The data collection at the university shows many irrelevant features that would not assist the research fulfil its aim. Therefore, another means of collecting more features became necessary and the best option is the use of questionnaire to acquire more information from students.

The researcher recognized that the federal owned university had information for only graduating students and that makes it difficult to locate the students to distribute questionnaires since they had finished their course and left the university premises. Hence, the questionnaires distributed were only to students in the state owned university.

To design the questionnaire, the researcher reviewed related literature to discover relevant features and listed out these features to help the research achieve its goal. From the features listed out, the researcher designed a questionnaire to help collect more features that are relevant from students.

The distribution of the questionnaire required the help of university staffs and three research assistants to identify, distribute, and retrieve questionnaires from the students identified. The research assistants helped in distributing the questionnaires while the staffs helped in communicating with and identifying low performing students. The researcher distributed a total amount of six hundred and fifty (650) questionnaires to low performing students in different faculties and levels in the university. Some students filled the questionnaires immediately while others opted to fill and return it back. The researcher and research assistants ensure students understood the questions and were willing to participate in the research, they also helped to clarify students in areas they were confused. The total number of correctly filled questionnaires retrieved by the researcher is four hundred and twenty seven (427).

Discussion

The data collected from the federal owned university contained little incomplete information; this could be because of the size of the data. The complete data the researcher received from this institution is about four hundred and seven six (476) records; from the complete dataset, one hundred and eleven (111) records are students with CGPA less than 3.0. The research focus is on low-performing students using 3.0 as the benchmark; therefore, one hundred and eleven (111) records is the dataset acquired from this institution. For the state owned university, incomplete and inaccurate data made up a huge number of the data collected. The student details collected from this university is about ten thousand four hundred and seventy two (10472) records; after manually inserting CGPA from the pdf files or hardcopy paper, the total number of students with CGPA is five thousand six hundred and thirty one (5631) records. Three thousand four hundred and eighty one (3481) records formed students with CGPA less than 3.0, which is the dataset collected from this university.

For the data collection process, the state owned university made the entire process tiresome, a student records department where all student records can be stored would have made the process a lot easily. The federal owned university with the examination office failed to gather all students' records but rather focused on final year level students alone; this in itself cannot produce a generalized model for the university.

During data collation, with the few records available in a spreadsheet format from the federal owned university, the researcher wasted no time and only had to sort student CGPA and extract students with CGPA less than 3.0 into another spreadsheet file. The process of collating the records obtained from the state owned university required a lot of time. First, the researcher arranged the student data collected from faculty officers into one spreadsheet file with each faculty having a sheet of its own to enable easy management of records, then added a new field called CGPA to all faculties. Next, the researcher manually keyed in the CGPA for each student record, which consumed a huge amount of time because of the size of the data. The researcher also spent time verifying that each CGPA keyed in tallies with the student data and that the figure is accurate. With the CGPA added to the dataset, the researcher sorted the CGPA field to extract all students with CGPA; this set of data for all faculties moved into a new spreadsheet file and sorted to get students with CGPA less than 3.0.

Data cleaning considered as an important step in the pre-processing stage of data mining handles the detection and removal of errors and discrepancies from data to improve its quality [13]. To ensure the dataset collected from the two universities are of high quality, the researcher handled the cleaning of data by ensuring that all records are complete and within required boundaries and also removed all inconsequential fields regarding the research purpose. Cleaning the data gotten from the federal owned university required less time and effort as the size of the data is small. The researcher noticed only two incomplete values in the dataset, one record with no value for sex of the student and the second record has no value for state of origin of the student. With only two missing values, the researcher opted to remove the two records because the dataset from the university had no field for student names, which would have assisted prediction of those values. The cleaning process for the state owned university proved very challenging because of the huge amount of incomplete data. There were several records with either no data or incomplete data, for example, about thirty-three (33) values were omitted for sex field, sixty-nine (69) values were omitted for date of birth field (with about twenty (20) incorrect values for year of birth), and eighty-three (83) values were omitted for state of origin field. These omitted values occurred randomly for all departments and some records contained two or more of these missing values. The data collected from the state owned university provided names of students; this assisted in predicting the sex and state of origin of some records. The missing values for date of birth used average date of birth values of students within the same department and level to predict the average date of birth for these records. The researcher also noted that about fifty-six (56) records had a CGPA of zero (0) and decided to remove these records because there is the possibility that these set of students registered for the courses but did not seat for the examinations; although it is also possible for students to fail all courses.

The researcher removed all irrelevant fields from the dataset for both universities, that is, fields that have the same values across all records. After cleaning the data, the researcher stored the clean dataset of in a spreadsheet file for mining. The clean dataset collected from the state owned university and the questionnaire retrieved from students merged together in one spreadsheet file formed the main dataset for mining.

Challenges

- Data collection from online data repositories offers huge amount of data that can be easily located and used [11]. However, the process of physically collecting stored data from any organization is more challenging and requires a lot of time. The challenges experienced during this entire research process are as follows:
- Delays: the amount of time spent in gathering data; from the time spent on waiting for feedback from stakeholders with regards to starting the actual data collection, the time spent on collecting the data, to the time spent on collating the data proved challenging.
- Locations of data: data was not available in one physical location in the state owned university; the researcher had to move from one faculty to another, even at the faculty level, student details are kept by different faculty officers while student result information are kept by either the heads of departments or departmental exam officers.
- Incomplete/Insufficient data: the amount of incomplete data from the state owned university reduced the amount of data by almost 50% and the federal owned university with data for only final year graduating students provided insufficient data.
- Data collation and cleaning: data collation and cleaning was tedious and required a great deal of attention. With this challenge, there is the possibility that the researcher might have made errors due to the large amount of data, although

the researcher verified severally to ensure these errors were minimal.

- Relevant data fields: the lack of relevant data fields at the end of cleaning the data available at the universities created cause for concern. However, with the help of the questionnaire, the research acquired more relevant data fields to aid the research in achieving its objective.
- The challenges experienced during the research were mainly challenges outside the control of the researcher, as the challenges required patience, time and a good attitude towards supporting staffs. In situations where the researcher could make adjustments to save time like communicating with two or three faculties concurrently, the researcher wasted no time.

Recommendations

For future researchers it is essential to map out their plans for the research and include delays in the planning, as it is inevitable since stakeholders might not be accessible at the same time. They are also encouraged to be patient and readily available in case of unexpected meetings. Researchers must have good knowledge of the data they need and be willingly to explain severally to different stakeholders at different times.

During collation and cleaning of data, researchers must endeavor to be careful to limit errors and look out for duplicate records. Researchers can plan the daily amount of records for collation and verify each record to improve its accuracy.

The research recommends better methods of storing data for the state owned, as the amount of incomplete data made up a huge amount of the dataset. The research also recommends that both institutions acquire and store more data, which would offer more insights about students for research purposes.

Finally, researchers can benefit from knowledge about data collection experience to assist them in future planning; therefore, this research encourage researchers in every disciple to include systematic documentation of observations, experiences and lessons learned during data collection into their research activities.

Conclusion

This research contributes to knowledge by describing the process of data gathering in a traditional learning environment that makes use of traditional methods to store student records. The lack of information management system makes gathering data within this environment demanding and time consuming. Researchers planning to conduct similar research must be patient, flexible in dealings with stakeholders and willing to spend a good amount of time. Further research is encouraged to offer more insight on data collection experience within this environment; this would highlight the inadequacies of the process and encourage these institutions to implement information management systems in their institutions. Although gathering data in this form is challenging and time consuming, it is also essential in order to support the goals of educational data mining in every learning environment.

References

1. Baker RS (2010) Data mining for education. Int Encyclopedia Edu 7:112-118.
2. Baker RS, Yacef K (2009) The state of educational data mining in 2009: A review and future visions. JEDM 1: 3-17.
3. Romero C, Ventura S (2010) Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40:601-618.
4. Merceron A (2015) Educational Data Mining/Learning Analytics: Methods, Tasks and Current Trends. In: DeLFI Workshops : 101-109.
5. Romero C, Ventura S (2013) Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3: 12-27.
6. Bruckman A (2006) Analysis of log file data to understand behavior and learning in an online community. In: The International Handbook of Virtual Learning Environments 1449-1465.
7. Ben Zadok G, Hershkovitz A, Mintz E, Nachmias R (2009) Examining online learning processes based on log files analysis: A case study. In: 5th International Conference on Multimedia and ICT in Education (m-ICTE'09).
8. Sife A, Lwoga E, Sanga C (2007) New technologies for teaching and learning: Challenges for higher learning institutions in developing countries. Int J of Edu and Dev Using ICT 3: 57-67.
9. Nsofor CC, Bello A, Umeh AE, Oboh CO (2015) The Future of Educational Technology in the 21st Century Nigeria: Changing Educational Landscape through Emerging Technologies. J Edu Pol Entrepreneurial Res 2: 28-37.
10. Romero C, Romero JR, Ventura S (2014) A survey on pre-processing educational data. Educational Data Mining: 29-64.
11. Siemens G, Baker RS (2012) Learning analytics and educational data mining: towards communication and collaboration. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge: 252-254.
12. Oyerinde OD, Chia PA (2017)Predicting students' academic performances: A learning analytics approach using multiple linear regression. Int J of Comp App 157: 37-44.
13. Rahm E, Do HH (2000) Data cleaning: Problems and current approaches. IEEE Data Eng Bull 23: 3-13.
14. Costa EB, Fonseca B, Santana MA, de Araújo FF, Rego J (2017) Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior 73: 247-256.
15. Mueen A, Zafar B, Manzoor U(2016) Modeling and predicting students' academic performance using data mining techniques. Int J Modern Edu Comp Sci 8: 36.
16. Guarin CE, Guzman EL, Gonzalez FA (2015) A model to predict low academic performance at a specific enrollment using data mining. IEEE Revista Iberoamericana de tecnologias del Aprendizaje 10: 119-125.
17. Ahmed AB, Elaraby IS (2014) Data Mining: A prediction for student's performance using classification method. World Journal of Computer Application and Technology, 2: 43-47.

18. Mythili MS, Shanavas AM (2014) An Analysis of students' performance using classification algorithms. IOSR-JCE 16: 63-9.

19. Hernandez Garcia A, Acquila Natale E, Chaparro Pelaez J, Conde MA (2018) Predicting teamwork group assessment using log data-based learning analytics. Comp in Human Behavior 89: 373-384.

20. Yassien NA, Helali RG, Mohomad SB(2017) Predicting student academic performance in KSA using data mining techniques. J Info Technol Software Eng 7: 187-191.