Academic Performance Analysis Framework for Higher Education by Applying Data Mining Techniques

Minimol Anil Job
Dept. of Information Technology and Computing
Arab Open University, Bahrain
m.aniljob@aou.org.bh

Jitendra Pandey
Dept. of Computing,
Middle East College, Muscat, Oman
jitendra@mec.edu.om

Abstract—Technology advancement and applications of various digital systems will help academic institutions to generate large amounts of data from different types of operational processes. Appropriate applications of data mining techniques in these large datasets can help the institutions to efficiently analyze hidden information and hidden patterns of data. Application of effective data mining techniques on educational dataset will help converting available data into knowledge. Then the mined knowledge can be filtered for decision-making. The application of appropriate data mining techniques and the results will support the academic institutions in their performance evaluation and improvement. Application of data mining techniques in e-learning systems will advance and enhance learning process in an educational institution. This paper presents a framework for predicting learner performance in higher education institutions by applying appropriate data mining techniques using data collected mainly from data collected from the Learning Management System (LMS) and other systems within the organization. This framework will help the institutions to make appropriate decisions about the learners' performance. Various clustering algorithms are suggested to apply in the data collected from various activities of LMS.

Keywords: Data mining, machine learning, clustering, educational dataset, learner performance, Artificial Intelligence

I. INTRODUCTION

Due to the advancement of technology and use of various digital systems to manage operational processes educational institutions are generating large amount of data. Educational institutions can efficiently utilize this data to make appropriate managerial decisions in learner performance evaluation. Effectual evaluations on collected dataset can be done to improve students' enhancement in academic activities, assess their learning performance, and evaluate their progression and retention rate. It can be used to evaluate as well as suggest mechanisms to enhance students' participation in different learning activities in an online learning environment. Also can be applied in forecasting the potentiality of students' performance in online assignments. Data Mining (DM) is a methodical process in which significant knowledge is extricated using various techniques from large data set. These dataset may contain structured and/or unstructured data. These datasets can be further applied with different DM techniques to obtain useful knowledge for future decisions. Some of the examples of DM techniques are clustering, classification, and association rule mining etc. DM techniques are efficient analytical tools in extracting meaningful knowledge from the available large data sets in academic institutions. Educational Data Mining (EDM) is defined as the application of data mining methods and tools for examining available data in academic institutions [11] [15]. Extracting learner's performance is a key area of EDM.

The LMS activities data can be collected from various registered courses by learners in an academic institution. The data collected from the LMS can be utilized properly and analyze to acquire profiles of students' behavior. One of the example is to monitor their performance in their online activities in the LMS. The performance indicators selected in this research are 'develop profiles of the learner', 'predict the potentiality of learner performance who can fail during an online assignment in LMS' and 'analyze the web data log files'.

This research paper has been organized into six sections. First section is the introduction, second section presents the related literature which describes various data mining techniques, EDM and applications. Section three describes the research methodology adopted in this research work. Section four depicts architectural framework of the student performance prediction system. The fifth section presents the framework model and hypotheses. The last section is the conclusions of the research and future work.

II. RELATED LITERATURE

978-1-5386-9346-9/19/$31.00 ©2019 IEEE

This section is based on the previous work done on the applications of data mining in education to measure the performance of learners. It is found that various papers have been published in this area of research. DM is a systematic process to extract useful knowledge from large amounts of data from stored in various storage repositories such as data warehouse [5]. DM techniques can be also applied in data generated from the e-learning systems to extract knowledge.

According to a research it has been concluded that DM techniques were used to analyze student achievement and student progression to the next level of study. The researchers applied a combination of quantitative analysis and qualitative using case study analysis. This combined approach of DM research was helpful in grasping about the various ways of actual implementation of DM techniques [21]. In a research paper, a methodology has been proposed by Yujie Zheng for application of clustering technique application in data mining. This methodology is suggested to improve the standard of higher education by finding data segmentation and pattern information [22]. There are many researchers conducted in the area of utilization of LMS data to evaluate learner performance by applying efficient DM techniques. Educational data mining uses many techniques such as decision trees, neural networks, k-nearest neighbor, naïve bayes, support vector machines and many others [19]. Another research paper proposed a model in which the researchers have described that by using different approaches based on clustering and sequential patterns techniques can be applied to find out new methods for improving the performance of students in academic institutions [18]. In a paper published by Jo & Kim, it has been presented that the enormous amount of data based on learner behavior in LMS can be collected and extracted and apply appropriate DM techniques to produce knowledge. This new mined data or the knowledge can be utilize in efficient way to improve learner's academic achievement [10]. As per the research results it has been observed that it is very challenging to predict student's learning achievement in institutions where they use an integration of traditional face-to-face and online methodologies as a medium of instruction [4].

The two broad categories of DM functionalities are descriptive and predictive analysis [11] [6]. The DM functionalities apply various methods and algorithms to discover knowledge and extract patterns of stored data [9] [11] [20]. The application of data mining techniques and tools to analyze educational dataset in institutions is defined as Educational Data Mining (EDM) [11] [7]. It is relatively important area in the DM research. In a research paper by Romero and Ventura, a survey of the application of data mining techniques to various educational systems is presented [11] [17]. In another paper by Romero et al. [11][17],[18] also based on EDM, discussed about the application of various DM techniques on collected data from various student activities using Moodle e-learning course management system.

From the various reviewed research papers it is concluded that determining and applying appropriate DM techniques on dataset on academic institutions will help the institutions in decision making. The data can be obtained from various systems used in academic institutions such as LMS.

III. RESEARCH METHODOLOGY

In this section, the researchers presents the three variables to be measured from the educational domain. The data availability and analysis using apocopate data mining techniques to measure these three variables.

The following variables are the outcomes to be determined using the data mining algorithms

The three performance indicators selected in this research are:

- Academic profile of students,
- Academic performance prediction and analysis, and
- Strategy planning based on Results and reports

As shown in figure 1, the process of data mining is divided into multiple stages which are discusses one by one.

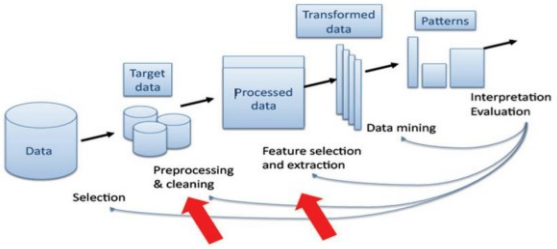

Fig 1: Data Mining Processes [8]

The first stage is data collection and processing, this stage is focused on data collection and execution. Depending on the research objectives data is collected from various department and specialization. Once data is collected, it needs to be stored properly so that it can be used further. This data will be further cleaned so as to eliminate any missing information.

978-1-5386-9346-9/19/$31.00 ©2019 IEEE

The second stage involves Transformation of data. This stage focuses on choosing write tool for the data analysis purpose. Like if we are using RapidMiner or SPSS then converting the data into the required format.

The third stage is focused on Discovery of required patterns by applying appropriate data mining techniques. Examples are clustering and classification [14].

The next stage is Knowledge discovery: in this step focusses on using extracted patterns for data analysis, extracting association rules.

The next step is Evaluation: knowledge from the data is extracted and how much efficient is this data is confirmed.

The last step is Action: After identifying strength or weakness of the data collected. This information can be used for various analysis.

Performance Indicator 1: To develop academic profiles of students

Data Collection method

The required data can be collected from various activities in LMS such as LMS activities data receive from log entries and results of forum activities. Figure-1 shows the LMS activities for a selected course for approximately two months. Figure-2 shows the forum activities on LMS for a selected course for approximately two months.

Data Mining Techniques

Two clustering methods are suggested for the collected data set, Hierarchical clustering and Non-Hierarchical Clustering methods.

- Application of Hierarchical clustering method:

Hierarchical clustering is a data mining method of cluster creation [3] [11]. This is a clustering analysis approach applied in a data set by partitioning it sequentially. The technique used in this method is constructing nested partitions layer by layer through grouping objects into a tree of clusters. In this method, there is no need to know the number of clusters in advance and uses a distance matrix as clustering criteria.

In this study, sample of the student data set, starts from the idea of 'students of a set' being more related to nearby students rather than far away students. Clusters will be formed from the data set by group of students based on their similar characteristics, in this case, similar LMS activities. The hierarchy of clusters will merge with each other at certain distances as per the algorithm.

- Application of Non-Hierarchical clustering method:

Non-hierarchical clustering uses a centroid based algorithm which takes to partitions of the data space into a structure [11]. The data space is a number of regions including subsets of similar data. In this study case of the student data set, the data mining method is used to generate groupings of a sample of students by partitioning it and producing a smaller set of non-overlapping clusters with no hierarchical relationships between them.

K-means algorithm [2] [11] can be applied in this data set. In order to function, k-means algorithms need seeds and the seeds can be determined randomly or by partial clustering. The students in the data set can be grouped on the basis of the seed points or the minimum distance between them. There are two types of movements, students will be transferred from one cluster to another or will be swapped with students from other clusters, starting from the initial classification, until no further improvement can be made in the data set.

Performance Indicator 2: Academic performance prediction and analysis

Data Collection Method

Data from SIS, LMS or any other information system used by the institution.

Data Mining Techniques

Various techniques of data mining like K-means clustering, Hierarchal clustering

Expectation Maximization

The expectation maximization data-mining algorithm uses 'parameter estimation' in probabilistic models with incomplete data. The expectation maximization algorithm is a natural generalization of maximum likelihood estimation to the incomplete data case. Expectation Maximization (EM) clustering algorithm [11] can be applied to discover student profiles from course evaluation data and for finding associations between subjects that was based on student performance.

Performance Indicator 3: Strategy planning based on Results and reports

The results and reports based on the data mining techniques applied to the available dataset. DM algorithms applies to the learner's data to produce outcomes. By making use of the rule engine a repository of knowledge can be created by storing relevant information, then rules and cases can be executed on any data. A framework model and hypotheses are given below to show the step by step DM process to reach into the results for the purpose of decision making.

IV. EDUCATION SYSTEM ARCHITECTURE USING DATA MINING AND ARTIFICIAL INTELLIGENCE

Graphical User Interface layer

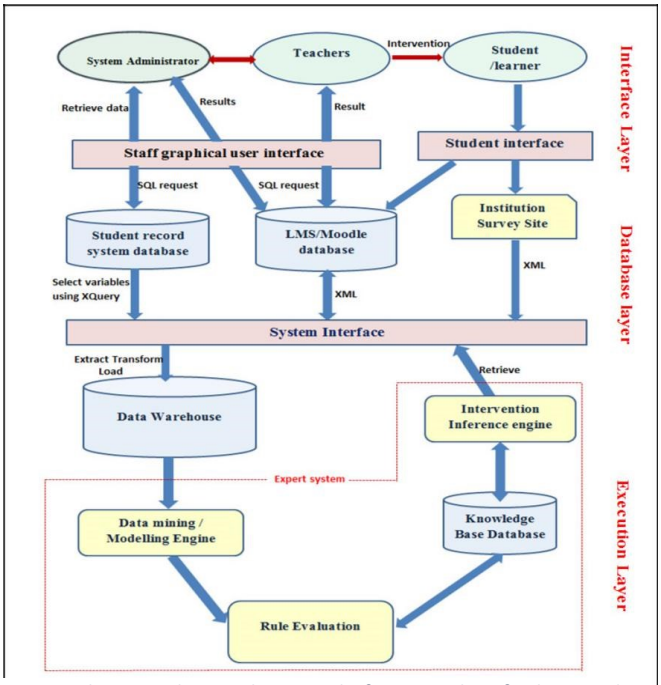

Fig 2 depicts the architectural framework of the student performance prediction system, which comprises of three different layers, a) the User Interface layer, b) the database system layer and c) execution or expert system layer.

This can also be referred to as the view layer. It hosts the Graphical User Interfaces (GUIs) of the framework. It is the layer that is presentable to the user and acts as the entry point to the system as well as provides necessary control and functionalities to the end users. It is divided into two categories based on the log-in interface, the staff graphical user interface and the student interface. With different levels of authentication, the staff and student can log in and carry out various activities.

Database systems layer

The database system layer provides access to the different databases available in higher education institution repository from where data abstraction for further analysis takes place. This is made up two categories of databases;

Firstly, the institutional databases which are made up Student Information System (SIS), Student Record System (SRS) and Learning Management System (Moodle) databases or any available information system in academic institutions [23][24]. Student demographic data such as age, gender, study mode, location, marital status as well as their online activities or actions such as number of time login, the frequency of login, total time spent online, assessment submission, forum activities and contribution can be extracted.

Secondly, the Student Psychosocial-Personality (SPP) database which manages students yearly psychosocial and personalities factors that are a not constant. Such SPP attributes include parental status, financial status, family responsibility, job workload, learning style, learning habit, parental academic level, parental support, academic environment, anxiety, student goal and interest, university support system, technology and social media impact.

Execution / Expert System layer

The execution or expert system layer consists of different units for modelling, evaluation and decision recommendation. It is faster and has low error rate than a human expert. The different units that made up the expert system are briefly explained below;

Datamining / Modelling Engine - This applies the selected data mining techniques such as characterization, classification, relationship mining, outliner analysis and clustering to the filtered educational/learners' data from the data warehouse. This will involve the application of the association mining rule to the training phases for generation of rules and patterns.

Rule Evaluation Engine – This uses logic and applies the set out rules, in a different form to the learner's data to produce outcomes. It makes use of declarative programming or conditional statement (IF and a THEN) to set out "what to do" and "how to do it" to produce the outcomes.

Knowledge-based database - By making use of the rule engine, it creates a repository of knowledge by storing relevant information, rules and cases that can be executed on any data.

Intervention and Inference Engine- This gets and combines information from the knowledge base to provide answers, suggestion, types and mode of intervention necessary for each student. It suggests and provides the necessary as well as a unique intervention strategy that the administrator and staff can use to support the student. Simply, it acts as a recommender of the personalized mode of intervention to the administrator and assigned staff about the learner. In addition, the inference engine also presents the dashboard, not just to the module tutor and the administrator, but also to the student.

V. FRAMEWORK MODEL AND HYPOTHESES

The Data Mining process in the proposed framework consists of four steps as shown in figure 3: The proposed framework comprises of various modules; data collection, Knowledge discovery, pre-processing of data, using data mining techniques in sequences steps. The process starts with classification and clustering of data specifically using various algorithms to identify the best result. Then to get result after visualizing these result to come up with the final decision which can be used by the institution management for the purpose of decision making.

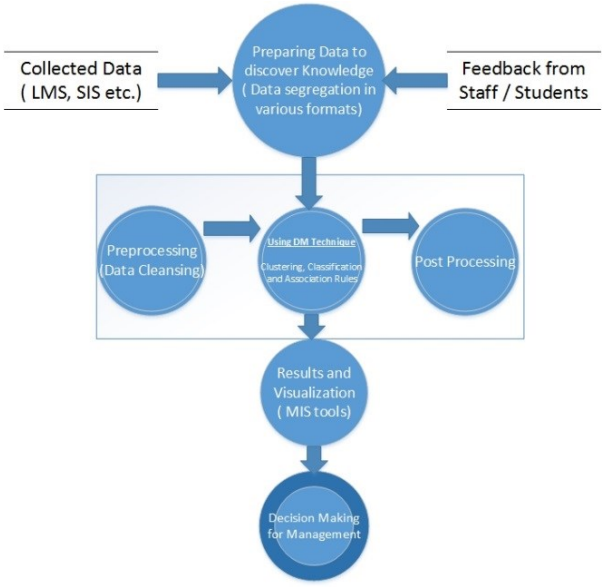

Fig 3: Research Framework Model

The proposed Framework in figure 3 describes how data collected from LMS can be utilized to apply the main DM techniques. The DM techniques suggested are classification, clustering, and association rule mining. Data collection is done through various information systems from the educational institution, such as SIS or Moodle or any other LMS. The next stage deals with the data cleansing mechanisms to ensure the quality of collected data as well as the identification of storage systems for preprocessing the collected and cleansed dataset. Next stage in the framework is Results and Visualization. After applying various DM techniques such as classification, clustering, and association rule mining in the academic dataset relevant knowledge is produced. Management Information system (MIS) tools can be applied for results and visualization of processed data. The results/information obtained from this stage can be utilized by the institution management for decision / strategy making purpose.

VI. CONCLUSION

With the increase use of technology, it is a good time for Higher education institutions to assess how new technologies can be used to assess various factors affecting the student performance during their educational career. The proposed framework in this paper comprises of various modules; data collection, Knowledge discovery, pre-processing of data, and data mining techniques in sequences steps. The process starts with classification and clustering of dataset specifically using various algorithms to identify the best result. Then visualizing these result to come up with the final decision which can be used for the purpose of decision making. The proposed Framework describes how data collected from LMS can be utilized to apply the main DM techniques. The DM techniques suggested are classification, clustering and association rule mining. Data collection through various information systems from the educational institution, such as SIS or Moodle or any other information systems. Data cleansing has to be done and storage for preprocessing to be identified. Applying algorithms and data mining techniques in the dataset will output results as knowledge. MIS tools can be applied for data visualization and the results can be utilized by the institutions' management for decision / strategy making purpose. DM and machine learning is a strong and effective tool which can help the universities and colleges to identify and allocate various resources. This can eventually help the management and staff to proactively accomplish student outcomes. This approach can additionally boost the efficiency of alumni growth. Having the capability to analyze huge data available in historical databases. Higher education institutions will be able to propose new educational models and ensure prediction with optimum accuracy of results based on the historical pattern of population groups. Higher educational institutions can plan and manage their

policies and strategies based on these decision making models, to efficiently discourse matters related to dissatisfaction of students resulting in switching between courses and ensuring retention, and strengthening of marketing and alumni relations. Researchers suggests Data collection from institutional information system, application of DM techniques and result analysis as the future work of this research.

VII. REFERENCES

[1] Baker, R.S.J.D. and Yacef, K. (2009) The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1, 3-16.

[2] Dan Pelleg and Andrew Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. ICML, 2000

[3] Douglas H. Fisher. Iterative optimization and simplification of hierarchical clustering. Journal of Artificial Intelligence Research, 4:147–179, 1996

[4] Garrison, D. R., & Vaughan, N. D. (2013). Institutional change and leadership associated with blended learning innovation: Two case studies. *The internet and higher education, 18*, 24-28.

[5] Gorunescu, Florin. Data Mining; Concepts, Models and Techniques, Springer-Verlag Berlin Heidelberg, 2011

[6] Han, J., Kamber, M. & Pei, J. (2011). 'Data Mining: Concepts and Techniques,' *Morgan Kaufmann,* San Francisco, CA.

[7] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, 443-491

[8] http://computation.llnl.gov/casc/sapphire/overview/overview.html

[9] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[10] Jo, I., & Kim, Y. (2013). Impact of learner's time management strategies on achievement in an e-learning environment: A learning analytics approach. *The Journal of Educational Information and Media, 19*(1), 83-107.

[11] Job, M. 2018. Data Mining Techniques Applying on Educational Dataset to Evaluate Learner Performance Using Cluster Analysis. *European Journal of Engineering Research and Science*. (European Open Access Publishing (EUROPA Publishing) 3, 11 (Nov. 2018), 25-31. DOI:https://doi.org/10.24018/ejers.2018.3.11.966

[12] Luan, Jing. Data mining and its applications in higher education, new directions for institutional research, No. 113, 2002, Springer

[13] Luan, Jing. Data mining and its applications in higher education, new directions for institutional research, No. 113, 2002, Springer

[14] M. Srinivas and C. Krishna Mohan, "Efficient Clustering Approach using Incremental and Hierarchical Clustering Methods", 2010 IEEE

[15] Renza Campagni, Donatella Merlini, Renzo Sprugnoli, Maria Cecilia Verri "Data mining models for student careers", at Science Direct Expert Systems withApplications,pp55085521,2015,www.elsevier.com.

[16] Rokach, Lior. "A survey of clustering algorithm," in O.Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, 2nd Edition, Springer, 2010

[17] Romero, C., Ventura, S. & Garcia, E. (2008). "Data Mining in Course Management Systems: Moodle Case Study and Tutorial," *Computers & Education*, 51(1), 368-384.

[18] Romero, Cristóbal, Sebastián Ventura, Pedro G. Espejo, and César Hervás. "Data Mining Algorithms to Classify Students." In EDM, pp. 8-17. 2008.

[19] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, **"**Data Mining Model for Higher Education System", Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010

[20] U. Fayadd, Piatesky, G. Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, AAAI Press / The MIT Press, Massachusetts Institute of Technology. ISBN 0–262 56097–6, 1996.

[21] Yeats, R, Reddy, P, Wheeler, A, Senior, C & Murray, J 2010, 'What difference a writing centre makes: a small scale study' *Education and Training*, 2010;vol. 52, no. 6-7, pp. 499-507

[22] Yujie Zheng, "Clustering Methods in Data Mining with its Application in Higher Education", International Conference on Education Technology and Computer, Vol. 43, 2012, IACSIT Press, Singapore

[23] Ahmed R, Pandey J. (2012), The Impact of Library Resources and Services on the Performance of Under Graduate Students at Middle East College of Information Technology (MECIT) Feb 16, 2012, WACAS, Muscat, Oman

[24] Joshi A, Pandey J. (2015), PLANNING EXTRA - CURRICULAR ACTIVITIES AT COLLEGE LEVEL TO ENHANCE STUDENT EXPERIENCE AND CAREER, publication date Nov 18, 2015 International Academy of Technology, Education and Development, Spain.