ISSN 2250-3153

Use of Data Mining Methodologies in Evaluating Educational Data

Thilina Ranbaduge

Department of Information Technology, University of Moratuwa, Srilanka

Abstract- Currently, online learning has gained a huge recognition within the higher education context and it has become a vital need in the current society to find such improvements, to increase the level of knowledge in people. In the modern society e-learning is recognized highly where it connects the students with the learning resources limitlessly. People have introduced various Learning Management Systems (LMS) in order to overcome the problem of managing large information sources and they are currently playing a remarkable role in e-learning environments. The lack of knowledge for employing learning methodologies in an accurate manner, has currently made these e-learning systems to face many problems. Even though LMSs enable the teachers to manage diverse educational materials in a much easier manner because of the differences in accessibility levels to the learning resources and study materials; it has become an unsolved problem to view the overall performance of each student in accordance with the behaviour of the student which is indicates the actual image of student learning capacity, on the course module. So it has become a massive challenge to cover the actual needs of the learners through the e-learning systems. Due to different learning patterns of students, it is becoming a vital need to understand the student performance, in a much more detailed manner.

Getting a proper understanding on a student overall performance which is based on the amount of information that he or she has gathered through the online resources, will help the teachers and the tutors to identify the different learning capacities of the students and will be able to provide the necessary guidance to improve their capabilities. To improve the learning capabilities of the students, the teachers and tutors should be capable of monitoring the overall performance of each student, separately and dynamically adjust their teaching methodologies on students and to take immediate decisions to improve learning of students. In this context, methodologies available in educational data mining can be used to extract knowledge from educational data sources to better understand students and the way they learn. This paper mainly focuses on the use of different data mining techniques upon the educational data to identify or discover the important knowledge on student learning which can be used to evaluate the students overall performances in the e-learning systems and identify and how these are been used to recognize different learning patterns of the students. Currently most of the techniques which are been used in each step in the data mining process contain its own advantages and disadvantages depending on the usage with the educational data which indicate the patterns of learning of students in many different forms and with various accuracy levels. However, based on these techniques, different models can be implemented to evaluate the performance of each student and it will be used to predict the overall performance that each student will be taking at the end of the course modules. Based on these results the teachers can provide the necessary guidance to the students who need more attention and also as an assistance to improve their capabilities on teaching and this will enable the knowledge producers to dynamically change the knowledge flows within the e-learning environments in a more effective and efficient manner.

Index Terms- Educational data mining, Evaluation models, Student data, Learning Management Systems

I. INTRODUCTION

In the current business environment, online learning has gained a major recognition within the higher education context and it is a vital need in the current society to find such improvements to increase the level of knowledge in people. The growth of Internet computing has enabled people to change the way of gathering knowledge and it has introduced a new drive way for distance education. In the modern society e-learning is recognized massively where it connects the students with learning resources limitlessly. During the past ten years educational institutions have integrated Information Technology with advancements in the communication fields for their educational programs to improve the level of teaching and the learning capacities of the students. Advancements in the Internet technologies have introduced various learning mechanisms for the students to gather more knowledge collaboratively and collectively, which has changed the way of gaining knowledge and teaching [2, 3].

II. E-LEARNING ENVIRONMENTS & LEARNING MANAGEMENT SYSTEMS

The growth of Internet computing has enabled the people to change the way of gathering knowledge and it has introduced a new driveway for distance education. In the modern society e-Learning has a massive recognition where it connects the students with the

learning sources limitlessly. Advancements in the Internet technologies introduce virtual learning environments to the students by enabling them to gather knowledge collaboratively and collectively. Moving away from traditional class room environments to the virtual Learning Management Systems (LMS) enable the teachers and the tutors to manage the diverse educational materials in a much easier manner [24].

Either commercial or open source, LMS are having a common purpose of providing the course materials for the students and open source implementations like moodle LMS are the most widely used virtual learning environments around the world. As shown in figure. 1, once a user logs in to the LMS he will be able to access all the courses for a particular semester he has registered and the LMS provides different e-learning activity modules (Calendar, emailing facility, news system, etc.) for the students to increase their interactions with the system.

Figure.1: LMS environments

The most important capability which is provided through the learning environments is monitoring and gathering information on each user's activities and behaviour within the system [7]. Each of the activities in which a particular student is actively participating indicates significant information about the student learning capacity. User behavioural information can be used for analysing purposes, where it can be used to identify different facts about the system and the current user's capabilities such that it will assist the knowledge producers to change the knowledge flow and the knowledge consumers to change the way of learning to improve their learning capacity [2].

III. WHY WE NEED EDUCATIONAL DATA MINING

Different advancements in the Internet and the telecommunication fields brought effective utilization of a large amount of knowledge which is diverse and distributed around the world. Sharing and manipulation of knowledge with distance interactivity in real time, where the teachers and the students are not seeing each other face to face is an enormous achievement that the society has gained by the use of IT technologies and the LMSs are used remarkably in order to overcome this problem of managing large information sources. Due to the lack of knowledge for employing learning methodologies in an accurate manner, currently e-learning systems are facing many problems [25, 28].

Virtual LMSs enable the teachers and tutors to manage diverse educational materials in a much easier manner. The LMSs provide many different grading mechanisms on the course learning resources that can be used by teachers to view the outline performance of each student and see the final marks that the student has gained in the given activity on the course [5, 10]. Due to the differences in accessibility levels for the learning resources and study materials; it has become an unsolved problem to view the overall performance of each student in accordance with the behaviour of the student which is indicated by the actual image of student learning capacity on the course module [25].

Learning and teaching in a virtual classroom environment depends on the way in which the information is flowing through the communication channels and the actual expectations of the users of the system. In modern e-learning environments, teachers and tutors are mostly dependent on the course module outline and these activities are arranged using the functionalities available in LMS according to the course module schedules. Due to the different levels of learning capabilities of the students these course activities are performed in dissimilar fashion by these students, which highlights that the knowledge level which they have gathered in each activity is different. Therefore applying the teaching principles or the teaching methodologies on all the students in the same manner will not fulfill the actual requirements of the students.

It has become a major challenge to cover the actual needs of the learners through the e-learning systems. Due to different learning patterns of students it has become vital to understand the student performance in a much more detailed manner without being concerned about the performance of each individual activity in the course module. A proper understanding of the students overall performance which is based on the amount of information that he or she has gathered through the online resources will help teachers and tutors to identify the different learning capacities of the students and will be able to provide the necessary guidance to the students to improve their capabilities since the main objective of an e-learning system is not to help the student to pass course module examinations but to help students to learn [13].

Learning is not only applied in the educational context, but it also can be applied for different organizations, where to use the information and learning processes effectively and efficiently to drive learning and development. Learning performance analysis reflects an organization's ability to execute performance driven learning effectively and efficiently to meet current strategic business needs as well as creating capability for the future. Performance driven learning ensures that investment in learning activities is always focused on vital performance elements and these elements are focused on performance factors that are critical to a learning scenario [27]. Based on the indications through performance analysis any organization can identify the level of the learning capacity that their stakeholders are currently in, and manage it appropriately as needed. This will ensure that the learning processes within the organizations always focuses on the predictive performance level and to achieve it by dynamically changing its driveways accordingly.

IV. USE OF DATA MINING METHODOLOGIES IN EDUCATIONAL DATA

Using analytics in the educational context to understand the student behaviour is an up-coming researching concept in the modern data mining arena. Becoming a new relative area of practice and research, different types of approaches and wide varieties of terms have been introduced. In the field of analytics, in the academy context many people have tried to come up with new computer supported interactive learning methodologies to identify the learning patterns of the students. Many people have researched on various computer aided software tools on collecting and analysing student data to build tools supporting in intelligent tutoring systems, games and learning simulation programs, to discover patterns and trends in a large content of student data to make new findings on hypotheses they made on student's learning [6, 22].

On most of the occasions, the available methodologies have come under similar conceptual and functional definitions, where most of the approaches were concerned on student learning data over the algorithms. But currently with such definitions, analytics are playing a major role in the higher education sector for the purpose of taking administrative decisions [27]. Currently most of the institutions are working with the administrative staff to manage the student enrolments in the study programs, to better utilize the resources, such as the budget, time and the staff among the study courses effectively and efficiently.

Researchers believe that the use of learning analytics in higher education will grow further with such importance, of identifying the student behaviour in the educational domain. Currently any types of institution, from a college to a university, academic analytics are being used to increase the financial and operational efficiency of the students. In order to cope with the high demands in the student's learning Luan indicated that many of the business critical questions are appearing parallel to higher education as well [27]. To address these issues in higher education many researches are currently implementing practices, focused on student retention, admission and operational efficiency.

According to the study conducted by the U.S. Department of Education [11] the most common reasons for student's to be dropping out of school are

- Lack of educational support - many students decided to drop out of high school due to lack of sufficient parental support and educational encouragement.
- Outside influences
- Special needs - students often drop out of high school because they require specific attention to a certain need, such as dyslexia or other learning disabilities
- Financial problems.

Out of the four mentioned above, the lack of educational support and the special needs reasons can be easily managed using the predictive analysis approach since student who are at risk of failing can be identified at an early stage by analysing their historical data of learning behaviour.

According to Natsu's report [23], it is mentioned that analytics can be used as a navigator for the education leaders to cut costs and improve teaching and learning in the institutions. She mentioned that the use of predictive analytics can be used for improving efficiencies to save money to enhance student achievement and the report included examples such as planning courses, recruiting and retaining college students, optimizing the scheduling of classrooms, and maximizing alumni donations.

According to the Cash, Dawicki, Sevick [8] the use of analytics, as an early warning system is a useful and effective tool and it should allow districts to achieve the following goals and objectives in their dropout prevention efforts:

Goal 1: Accurately define and uncover students' problems and needs
Goal 2: Successfully identify interventions and improvement strategies
Goal 3: Effectively target and initiate programs and reforms
Goal 4: Truthfully monitor ongoing efforts and progress with 'at-risk' students

According to them, in order to achieve these goals either Business Intelligence or Predictive Analytics can be used. Business intelligence model and tools are more focused on analysing historical and current data in order to provide a look at operations or conditions at a given period of time while predictive analytics approach attempts to incorporate historical data into statistical models in order to make predictions about future events or outcomes.

A. The Knowledge Discovery Process

Application of data mining with educational data sources is a recent research domain which has gained a larger recognition among the society. Proper understanding on the learning behaviour of the students helps the educational sources to manage their educational programs, to a much more improved level, which can increase the learning capabilities of the student, who followed their educational programs.

Academic analytics or Learning analytics is a wide term used recently to describe the use of data mining in educational data sources [27]. Researchers have used various data mining methodologies in different ways to understand or identify the learning models and learning patterns of the students. Various learning management systems can be used for providing the study programs for the students who can be connected in open distance mode or in a much more hybrid manner. But applying the same teaching principle on all the students in the same way will not cope with the ultimate goal of any learning management systems available, which is to help the students to learn rather helping them to pass.

But understanding the nature of the learning in each of the students requires a huge effort such that the researching of educational information need to be followed by a stepwise process to reach the final discoveries. Like in any data mining research, some steps can be given as the basic steps that need to be followed in the knowledge discovery process which is shown in figure 2. The same principles can also be applied in the learning analytics processes where the data collecting step will be applied to different educational data sources to aggregate necessary data for the analysis and data pre-processing steps will be applied to arrange the data into a proper arrangement which can be applied on the mining algorithms to discover the hidden information and patterns [3, 5].

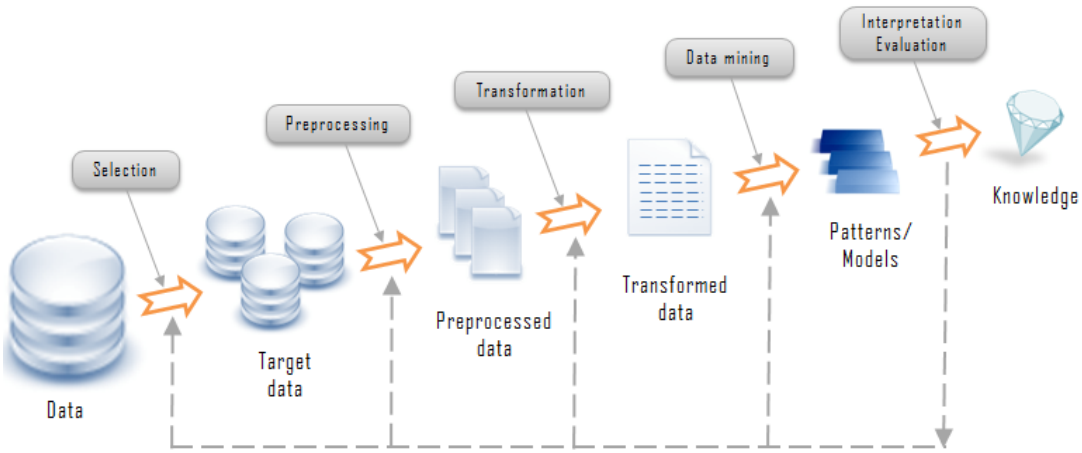

Figure. 2. Knowledge Discovery Process (Reference: http://www.rithme.eu/img/KDprocess.png)

B. Institutional Implementations on Educational Data Mining

The importance of using analytics in an academic context has been widely recognized among the institutions. Most of the institutions or universities are using the learning analytics approaches to improve their student enrolments in the course modules which are offered. Some researchers have tried to come up with complex models which are based on examination scores, course works and other related information, to discover the patterns in student's enrolments in educational programs [24, 25].

The main advantage which they have achieved through such analytics approaches is to implement an improved management process on the efficient use of limited admission budgets, time and the available staff which can be viewed as a statistical analysis approach on diverse institutional data sources. To create the required level of intelligence in an educational context many approaches

have been implemented over the course management systems to improve teaching, learning and success of the students which gives an early prediction on which students are in academic difficulty and by allowing the faculty and advisors to customize learning paths or provide personalized instruction to specific learning needs.

Baylon University, which is a pioneer in higher educational analytics, created an Enrolment Predictive Model as a supportive tool for the prospective student admissions [22]. Moving from traditional admission strategy to the predictive model they analysed factors which provides various aspects in the student. Most of these variables they were concerned are based on the student's motivation on attendance, extracurricular activities and score values attained at different levels of examinations. Scores generated by the predictive model are considered by the admissions staff to identify those students most likely to be admitted to the university and they achieved the advantage of allocation necessary for human and budgetary resources appropriately.

Purdue University has developed a prediction model which extracts data from the Course Management System (CMS) and predicts which students may be at a risk in academic work [22]. The main purpose of this system is to provide proactive involvement in students learning which is mainly focused on student academic success which is the result of the student's capacity on standardized test scores and other similar information and the student's effort and motivation which is measured by the participation within the CMS. The prediction model is been prepared for two user groups of freshman and overall student population. Using factor analysis and logistic regression mechanisms these models were tried to predict the student success in a given course module.

In year 2010 Purdue University implemented a new analytic tool known as Course signal [24] to increase the student success in the classrooms. This system is capable of providing early warning about the students who are not academically performing well in the classroom. Both the student and the teachers can benefit from the system so that the real time feedbacks and necessary communications will be arranged for the students to find the required resources to improve their knowledge levels in academic subjects.

A research which has been taken in Northern Arizona University (NAU) tried to model connectivity between the resource utilization, level of risk for the student and their outcomes to create a proactive involvement in their academic achievements [22]. The researchers found that the advising and using of resources can be efficiently done on the student's who are at high risk level in academic works through academic recommendation and career sessions. They also indicated that having an intrusive advising mechanism on students will help them to understand the effective use of timing, content communication and the careful planning on the subject matters.

The University of Alabama introduced a new predictive model which can be implemented to predict and improve the student retention at the university with analytics approaches [22]. The researchers have developed a sophisticated model with attributes on student data by using statistical techniques such as logistic regression, decision trees and neural network mechanisms. Based on the information which has been discovered by the model, the faculty and the academic advisors have given instructions for the students to select the necessary course modules.

C. Use of Learning Analytics in E-Learning Environments

LMSs have started its process as assistant to a much more crucial role in higher education. Most of the organizations, Institutions and Universities currently use many different types of Learning Management Systems to provide their study programs for the students. The growth of the Learning Management Systems have been a challenge between two categories of commercial or the open source where these two are competing with each other for providing a better service to the users [29, 30].

The major open source learning management systems those currently dominating the educational space can be specified as Moodle and Sakai [5]. The major factor behind the demand for this open source LMSs is because of their cost saving and more control. With the available budgetary constraints people tend to move to open source LMSs due to the fact that any function found in the commercial tools is available with the open source LMS as well. As mentioned in the Sakai Project web Site the main objective of the Sakai is to design, build and deploy a new Collaboration and Learning Environment for better higher education. Sakai was initially started by the University of Michigan and Indiana University where both put their efforts to enhance the functionalities. After the grants given by the Mellon Foundation, MIT and Stanford joined in and formed the Sakai Project.

It is found that from an administrator's perspective, the Sakai system is relatively easy to install and set up compared with the many other open source projects in their early stages of development which does not seem to be easy in its installation process. Even some commercial tool users commented that the Sakai user interface experience is very much familiar and easy to manage. Moodle is a course management system which is a free and Open Source software package designed using sound principles to help educators create effective online learning among the students.

Lauria and Joshua [5] have tried to implement a predictive model within the Sakai for predicting the performance of the students and to take the decisions for making corrective actions. In their research they came up with a methodology which contains six phases

on the knowledge discovery process which are data collection, data reduction, rescale and transformation, partitioning the data, build the models and finally to evaluate the models and choose an appropriate model for the predictions.

In any educational institution the learning capacities and the learning capabilities of the students vary on different levels due to the way they arrange their learning behaviour. Due to this learning nature applying the same teaching principle on each and every student in the same manner does not provide the required level of knowledge in students. This gives an indication to the teachers why many students are dropping out from schools or educational institutions.

In the year 2009, Massachusetts Department of Elementary and Secondary Education started a project known as the Dropout Prevention Planning Project to implement an approach to understand the different factors on student's dropout incidents [8]. They implemented a Student Information Management System which contains information about the student within the State and they initiate an index known as the Early Warning Indicator Index for measure the risk of dropping out in each student.

According to the Massachusetts Department of Elementary and Secondary Education, currently there are several early warning systems available, which are using predictive analysis approaches to analyze the students. Microsoft SIGMA is an early warning system which capitalizes on its Education Analytics Platform (EAP) to provide a new data-based approach to managing students who are at-risk and it is known as the Student Individualized Growth Model and Assessment [8, 11]. Mizuni's Data Warehouse and Dashboard Suite [8] is a transactional and aggregation data store, for managing and analyzing data and offers education stakeholders insight, into student performance, by monitoring key indicators to increase student achievement. VERSI-FIT also based on Microsoft has also developed its own early warning system based upon the Education Analytics Platform which is known as the Edvantage At-Risk Early Warning System and Credit Recovery System [8].

D. Collecting Educational Data for Analysing

Understanding the way these students are learning will help the teachers and the academics to change the teaching methodologies they used, to cope with the requirements of the students. As the academic analytics suggest, the knowledge can be created about the students, by applying the statistical analysis and predictive modeling on educational data sources. For building these analysis processes what is mainly used is the continued streams of data which are created within the learning management systems with the data mining techniques and which can be used as a decision making tool for teachers. Before applying the analytical modeling it is a must to acquire proper information out from the student data since not all information available in the data sources is important for the analysing [7].

The main advantage which in the given context is that most of the data which are available in the data sources are labeled and they contain information about the student characteristics as well as the course management events where the modeling process can be applied in many ways. The use of LMS systems in researches assist the researchers in gathering of necessary data which are collected for different course modules. Most of the LMS systems maintain such student data using log files where each log contains information about the activities each student has performed during the course schedule [6, 25, 29]. Using the activity information with the demographic characteristics of students such as age, sex, residence etc. can be used with the mining algorithms to understand how students are motivated on learning and how the performance of each student varies with their learning patterns.

Lauria and Joshua [5] indicated the importance of gathering the student log data for the analysis process and discovering patterns in the student data. In their research they collected the student data from diverse sources and followed several pre-processing steps to handle the missing value, outliers and incomplete records. All the student academic performance data which were logged in a Sakai implementation named iLearn were aggregated with the student demographic data such as age, gender, race, SAT scores, GPA and with the course enrolment data such as course name, course subject, number of students in the course to produce consolidated records per course and student. In order to remove the variations in different course contents all the data were collected as ratio values rather than as an absolute value.

One of the main reasons that applying learning analytics in educational data sources is challengeable is because the dataset becomes very small compared with the other applications around in the given context. Even though the number of student information which is contained in a database is huge, most of these are dynamic and contain many variations among them. Diego Garcia Saiz and Marta Zorrilla [10] found it difficult to collect the required data which made them to use the data for the past three academic years with an average student enrolment of 70 per year, for a specific course module, to build the necessary dataset for the analysis.

For all of these student instances they considered attributes with mean values such as total time spent, number of sessions carried out, number of sessions per week, average time spent per week and average time per session. In S. Anupama Kumar and M Vijayalakshmi [12] research they have tried to predict the student overall performance based on their internal assessments in the learning environment and they considered five course modules which were offered in a semester and overall student count for the selected analysis were about 117.

Even though the datasets are small compared to the other data mining domains, still it can be proven that sensitive and interesting information can also be generated regardless to the size of the dataset. T. Hadzilacos, Dimitris Kalles, Christos Pierrakeas and Michalis Xenos [14] indicated that even in a complex educational environment, sensitive learning patterns can be identified, applying the machine learning techniques on a small set of student data. In their research they focused on applying various algorithms on demographic data and student course data to discover the relationship of tutors with their students.

Other than the size of the dataset attributes or the collection of independent performance factors it is also highly concerned with the educational data mining researching context. In educational systems like Learning Management Systems, Students' academic performance depends on diverse factors like personal, socio-economic, psychological and other environmental variables [28]. Each of these factors can affect the student overall performance in different weights. Based on the level how each of these factors is appearing in the student education, several learning patterns can be identified on each of these students. Based on these learning patterns analysis models can be implemented such that they include all these variables for the effective prediction of the performance of the students. The prediction of student performance with high accuracy is beneficial to identify the students with low academic achievements which enable the educators to assist those students individually.

S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas [18] have tried to apply data mining methodologies on educational data to limit student dropout in university-level distance learning where they argued that the dropout can be caused by professional, academic, health, family and personal reasons and varies depending on the education system adopted by the institution providing distance learning, as well as the selected subject of studies.

They based their research on a course module which was offered in Hellenic Open University which based their educational programs, mainly on distance mode. They built a data set of 365 student instances and based on the data the attributes were divided into two groups which were the 'Curriculum-based' group and the 'Students' performance' group. The 'Curriculum-based' group represented attributes of students' sex, age, marital status, number of children and occupation and the group represented attributes concerning students' marks on the first two written assignments and their presence or absence, in the first two face-to-face meetings.

In the M. Ramaswami and R. Bhaskaran [15] research, they argued that the student performance could depend on diversified factors such as demographic, academic, psychological, socio-economic and other environmental factors. Based on each of these factors they constructed their analysis process to identify the different performance factors of the students and they obtained highly influencing predictive variables through feature selection technique to evaluate the academic achievement of students.

The data collection process which is in T. Hadzilacos, Dimitris Kalles, Christos Pierrakeas and Michalis Xenos [14] research, they collected student data under two categories of attributes, which are Demographic attributes and Performance attributes. The Demographic attributes were collected by concerning students' sex, age, marital status, number of children and occupation and Performance attributes represents attributes which were collected from tutors' records concerning students' marks on the written assignments and their presence or absence in face-to-face meetings. Other than the above it was found that there exist some obvious and some less obvious attributes that demonstrate a strong correlation with student performance where some give the higher importance in consideration.

Other than the demographic and performance student factors some data can also be collected through the students activities which they have performed within the learning management system. The data related to the discussions, quizzes and messaging can highlight significant information on student communications between their other peer students, teachers and tutor and provide information about how they have gathered information through communications channels which are provided by the learning environment.

J. Mamcenko, I. Sileikiene, J. Lieponiene, R. Kulvietiene [26] have tried to collect data on questions given to the student and the answers given by the students with the amount of time that each student has spent on each question, to construct the dataset and applied the collected dataset with several data mining methodologies, to discover the different learning patterns of each student which they focused on a programming course examination and the ways that the students have answered.

C. Romero, S. Ventura, P. G. Espejo, and C. Hervás [2] have tried to use different classification approaches on the student activity data, to compare the applicability on data mining techniques for classifying the students into groups and to predict the final marks obtained in the course modules. For constructing the dataset they used the activity information from the database in the moodle environment and they extracted the information on moodle activities and the final marks the students have achieved for 7 course modules.

E. Data Pre-Processing Techniques in Educational Data Mining

In any knowledge discovery process pre-processing steps that apply on the collected data gain, a huge consideration where many mechanisms have been introduced by different researchers. The main objective of applying these pre-processing steps on the

aggregated data is to increase the accuracy levels and to improve the quality of the data since the collected data can contain noises or improper data for applying on the data mining algorithms [6].

In an academic context the LMS systems contain much larger data sources which need to be included together to create the dataset [28]. When collecting the required data from diverse sources, extracted information can contain missing value, outliers and incomplete records. Applying the data mining techniques on the dataset without proper preparation will provide unexpected and misleading results as outcomes. Lauría and Joshua Baron [5] highlighted the importance of applying pre-processing steps on educational data. After the data collection process they followed some steps to reduce the dimensions on the available data. In order to maintain a proper level of query accuracy and efficiency the number of variables and parameters required for the estimation were selected properly and unnecessary features were removed.

After the necessary data was selected the transformation and rescaling phase was carried out to make sure that all the attribute data were formatted according to the requirement of the data mining algorithms they used. After the data was converted or transformed the partitioning step was used to divide the data into several groups. They carried out this partitioning process on the data set to make sure that the required amount of data is available for the training of the data model and for the validation with a testing step.

Outliers and noisy data in the data set can affect the accuracy levels and the quality of the outcomes in the algorithms. To achieve more improved and accurate results D. García-Saiz and M. Zorrilla[10] built an algorithm known as 'meta-algorithm' to pre-process the dataset and eliminate these outliers. In this algorithm, the data instances with the highest values for the most significant attribute was removed such that these instances can be considered as the outliers in the statistical sense which can improve the results by 20% and which would make a huge advantage when the data set is larger in size and provide with better quality results for the users.

Other than the noises and outliers in the dataset, selecting the most appropriate set of attributes for the data mining algorithms, improve the accuracy levels furthermore. S. Kotsiantis, C. Pierrakeas, and P. Pintelas [18] indicated that there exists some obvious and some less obvious attributes that demonstrate a strong correlation with student performance where some give the higher importance in consideration. As the author suggested by the research, identifying the correlated attributes in the given data set can be used to improve the results of the algorithms and to reduce of dimensions as needed for the algorithm for efficient processing.

In J. L. Hung and K. Zhang [9] researched on activity patterns and making predictions with data mining techniques in online teaching and the data pre-processing phase of the research the data was cleaned by removing all useless, irregular, and missing data from the original LMS common log files and after the initial pre-processing, a session filter was applied to the reduced log file for feature extractions. The purpose of the filter was to aggregate all user requests within a session into a single set of variables. Feature extractions filtered out the following primary variables: user identifier, session identifier, session start date and time, session end date and time, user's hit count, and session duration in minutes and based on these derived variables (duration and frequency of data of each student) were extracted through calculating or accumulating primary variable data on a daily and weekly basis.

F. Use of Analytic Algorithms in Evaluating of Educational Data

According to the requirements in the problems domain, discovering a model to evaluate the students' performance levels, created a key researching area, where different researchers have attempted to find the accurate and possible models [27]. Due to the nature of educational data sources and student behavioural patterns the same performance model cannot be applied to cover all the problem scenarios. Based on the different approaches available on the educational data mining or learning analytics, different areas which need to be considered, can be identified.

The k-nearest neighbour (KNN) data mining method is a classical prediction method among the machine learning techniques available in data mining [3]. It has been widely used due to its simplicity and adaptability in predicting many different types of data. The main advantage of using KNN in prediction processes is that the KNN is a lazy method which does not require a model to represent the statistics and distribution of the original training data. Rather it can be applied on the actual instances of the training data. Even though the KNN is a simple predictive algorithm which can rely on and it does not make any assumption about the prior probabilities of the training data. Also the KNN is satisfactorily used on the situations when the data set is included with noisy and incomplete data.

Due to the advantages and the simplicity of the algorithm T. Tanner and H. Toivonen [21] have tried to implement a model using the k-nearest neighbour data mining algorithm to identify the students who are at high risk of failing in a specific course. By this research they suggested that good results in predicting final scores indicate that students with learning problems can be found reliably. What they have been using on the student data is to make prediction on the performance of a given student based on the similarity to all instances in the training set and find the k most similar objects in the data set. This similarity is calculated by using a Euclidean distance between the features of the test subject and the corresponding features of each instant in the training set.

In their research they showed that KNN can produce predictions accurately for the final scores even after the first lesson. Another interesting result they found is that, in any skill-based courses early tests on skills can be used as the predictors for the final scores and they suggested that predicting final scores for the courses can be used to identify the students with learning problems and can be used directly to implement as early warning features for the teachers so that the students can be alerted if they are likely to fail the final tests. Based on the information or features they used for the experiment with the KNN algorithm they suggest that the KNN method could be just as effective in other LMSs such as Moodle where only a single lesson score is available for student assessment and especially other skill-based courses could be a good selection for the KNN method.

In order to make the student modeling process much easier Diego Garcia Saiz and Marta Zorrilla [10] have researched on applying different classification techniques on the student data to predict their performances. In their research they tried to implement a tool known as Elearning Web Miner (EIWM) to discovering how the students are behaving and progressing in the courses which is very helpful for the tutors to identify the students who need more attention among a larger set of students. With the intention of analysing and choosing best classification algorithms for educational datasets they analysed four of the most common machine learning techniques, namely Rule-based algorithms, Decision Trees, Bayesian classifiers and Instance-based learner classifiers which mainly were OneR, J48, Naive Bayes, BayesNet TAN and NNge.

In the evaluation process they found that Bayes algorithms perform better in accuracy and is comparable to J48 algorithm although it is worse at predicting than Naive Bayes which is the best in this aspect. They also observed that NNge improves its performance in this dataset although the great number of rules which it offers as output makes it less interpretative for instructors than the rest of the models. Finally they conclude that Bayes Networks are suitable for small datasets in performing better than the Naive Bayes when the sample is smaller. As a consequence of the fact that BayesNet TAN model is more difficult to interpret for a non-expert user and J48 is similar in accuracy to it.

S. Kotsiantis, C. Pierrakeas, and P. Pintelas [13] have suggested an approach which has used machine learning algorithms with the LMS data to prevent, student dropouts in university distance education. In their research they used five different algorithms to study student data and they found that these algorithms can be used more appropriately to predict the student dropouts in study programs. In this research they used most common machine learning techniques which are Decision Trees, Bayesian Nets, Perceptron-based Learning, Instance-Based Learning and Rule-learning [3]. In the evaluation of the algorithms they found that there was no statistically significant difference between algorithms, but it showed that the Naive Bayes algorithm and the RIPPER had the best accuracy than the others. Among the Naive Bayes algorithm and the RIPPER, Naive Bayes has the advantage short computational time requirement and importantly Naive Bayes classifier can use data with missing values as inputs, whereas RIPPER cannot work with which gives an indication that the Naive Bayes is the most appropriate learning algorithm to be used for the construction of a software support tool in Learning Management Systems.

This research was further enhanced by S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas [16] by applying six machine learning techniques which are Decision Trees, Neural Networks (NN), Naive Bayes algorithm, Instance-Based Learning Algorithms, Logistic Regression and Support Vector Machines. Based on these six algorithms they found that Naive Bayes algorithm and the NN algorithm had the best accuracy with the given data sets. However they mentioned that the differences were generally small and because they were only based on one course module and it may be possible that the ranking in another data set of the same domain is different. Also they concluded that Naive Bayes has the short training time and an effective communicated way of predicting and a small programming cost than the other algorithms.

A research which was completed by C. Romero, S. Ventura, P. G. Espejo, and C. Hervás[2] have tried to used different classification approaches on the student data to compare the applicability on data mining techniques for classifying the students into groups. In their research they used a framework which is known as Knowledge Extraction based on Evolutionary Learning (KEEL) which is an open source framework for building data mining models. For this research they used 25 classification algorithms which are based on Statistical classification, decision tree, rule Induction, a genetic algorithm using real-valued genes, fuzzy rule induction, neural Networks. According to their research they found that models obtained by using categorical data are more comprehensible than when using numerical data because categorical values are easier for a teacher to interpret than precise magnitudes and ranges.

Also decision trees are considered as easily understandable models because a reasoning process can be given for each conclusion. But a tree obtained with large nodes and leaves are less comprehensible. Rule induction algorithms are also considered to produce comprehensible models because they discover a set of IF-THEN classification rules that are a high level knowledge representation and can be used directly for decision making. Fuzzy rule algorithms obtain IF-THEN rules that use linguistic terms that make them more interpretable by humans and these rules are very intuitive and easily understood by problem-domain experts like teachers. Finally the statistical methods and neural networks are deemed to be less suitable for data mining purposes due to the lack of comprehensibility even they attain very good accuracy rates but very difficult for people to understand.

S. Anupama Kumar and M Vijayalakshmi [12] have tried to predict the student overall performance based on their internal assessments based on decision tree approaches. The algorithms they used for this research is J48 and ID3 decision tree algorithms. According to their discussion the accuracy level of the J48 was higher than the ID3 algorithm since it has predicted more correct prediction results than the ID3 algorithm. Based on the different accuracy levels on each of the decision trees they created, they concluded that classification techniques can be applied on educational data for predicting the student's outcome and improve their results and the efficiency of various decision tree algorithms can be analysed based on their accuracy and time taken to derive the tree. Finally they argue that the application of data mining brings a lot of advantages in higher learning institutions so that these techniques can be applied in the areas of education to optimize the resource allocations as needed with the student learning capacity.

Decision tree approach was further examined by M. Ramaswami and R. Bhaskaran [15] research where they have argued that the student performance could depend on diversified factors such as demographic, academic, psychological, socio-economic and based on these factors they constructed a CHAID prediction model with highly influencing predictive variables obtained through feature selection techniques to evaluate the academic achievement of students. In their research before constructing the CHAID model they used feature selection techniques in reduction of computation time and enhance the predictive accuracy of the model.

In their analysis, they found that the overall prediction accuracy of CHIAD prediction model was higher compared to the other classification mechanism on categorical attributes and they suggested that even though CHAID model is capable of handling small and unbalanced data set, it could be worked out effectively with more predictive accuracy which can be further improved by applying some principle pre-processing techniques.

In the research of Dimitris Kalles,Christos Pierrakeas [17] they tried to use a genetic algorithm and decision tree based classification on student data to understand the different learning capacities of the students. In their research they based the applicability of these algorithms on different sets of students under different course modules. In this research they mainly used the genetic algorithm based decision tree implementation of Genetic Algorithm Tree (GATREE) which is built using the Genetic Algorithm Library (GALIB) library.

In this research GATREE system and experimented with to 150 generations and up to 150 members per generation. They observed that GATREE induced trees provide good accuracy estimation, even without the cross-validation testing phase. Their initial findings suggested that when compared to conventional decision-tree classifiers this approach produces significantly more accurate trees. However it was noted that GATREE has been generating closer estimations even with the quantized formats which gives an indication that GATREE can produce quality results even in the presence of noise.

In Behrouz Minaei-Bidgoli, Deborah A. Kashy, Gerd Kortemeyer, William F. Punch [20] research, they highlighted the importance of using Genetic Algorithms not as a direct classifier on the data but as an optimization tool for resetting the parameters in other classifiers. In this paper they focused on using a Genetic Algorithm to optimize a combination of classifiers. They used Genetic Algorithm Tool Box (GAToolBox) for MATLAB to implement a Genetic Algorithm to optimize classification performance and to find a population of best weights for every feature vector which minimize the classification error rate.

The same idea was further enhanced by the research of Behrouz Minaei-Bidgoli, Gerd Kortemeyer, William F. Punch [19], where they have used the Genetic Algorithms to find a population of best weights for every feature vector from the attribute set which can minimize the classification error rate and it was found that the genetic algorithm for weighting the features improved the prediction accuracy by 10% for the other classifiers used in the research.

Most of the previous researches which were carried out, considered each classification approaches individually for building the performance models and different algorithms are having varying levels of accuracy levels. But in Behrouz Minaei-Bidgoli, Gerd Kortemeyer, William F. Punch [19] research, they restricted their study to four different classifiers which were Quadratic Bayesian classifier, 1-nearest neighbour (1-NN), k-nearest neighbour (k-NN), Parzen-window and try to model the performance of the students by combining the classification algorithms together. As the conclusions they mentioned that a combination of multiple classifiers leads to a significant accuracy improvement in the given data sets.

Other than the classification approaches, clustering is a data mining technique which can be used to identify the hidden patterns in a given dataset [3]. Jui-Long Hung and Ke Zhang [9] tried to use a clustering technique to classify students based on their shared characteristics. They used K-Means clustering mechanism to identify the student clusters based on the behaviours they showed within the LMS environment and Sequential association rules were applied to discover the daily learning patterns of the students in the LMS. Finally they used decision tree algorithm to build the predictive model on the students. According to the predictive model, the frequency of accessing course materials was the most important variable for performance prediction in this study. Also this study concludes that when students participated more actively that is having a higher value on frequency of accessing course materials, number of messages posted, number of messages read, and frequency of synchronous discussions attended they performed better than the others academically.

Finally in this research paper the authors suggested that instructors would be able to get a quick view of basic learning data, such as login date, frequency, pages visited, etc. However, no functions or features are currently available to help instructors identify learners' individual or group learning patterns, or to identify successful or less successful learning behavioural patterns, or to identify the predictive learning behaviours or to help identify necessary facilitation needs. Therefore, the researchers of this study strongly suggest that LMS developers should integrate data mining tools to facilitate effective online teaching and learning.

The clustering mechanisms were further examined by the research of J. Mamcenko, I. Sileikiene, J. Lieponiene, R. Kulvietiene [26] where they have tried to apply Kohonen algorithm which is based on a self-organizing map (SOM) or self-organizing feature map (SOFM) that is a type of artificial neural network (ANN). According to their research it can be concluded that the clustering mechanism they used can be used to discover the statistical information about the student behaviour and the learning patterns.

According to the Hershkovitz and Rafi Nachmias [4] web usage mining is another information source which can be used to analyse the whole learning process and to examine the activity of a large group of learners, in order to develop a log-based motivation measure on LMS environments. Finally F. Castro, A. Vellido, À. Nebot, and F. Mugica [1] have researched on applying various data mining mechanisms to detect and understand the irregular learning behaviour of the students and the adaptability on LMS environment on student's requirements and capacity.

Finally B. Minaei-Bidgoli, G. Kortemeyer, and W. F. Punch [19] suggested that use of these algorithms as tools can be used to identify those students who are at risk in very large classes and it will help the instructors to provide appropriate advising in a more effective manner. Also J. L. Hung and K. Zhang [9] indicated the importance of implementing such tools which can be used to identify learners' individual or group learning patterns or to identify successful or less successful learning behavioral patterns, or to identify the predictive learning behavior or to help identify necessary facilitation needs. LMS developers should integrate data mining tools to facilitate effective online teaching and learning.

V. CONCLUSION

During the past decades advancements in the Internet technologies introduced various learning mechanisms to students to gather more knowledge collaboratively and collectively. These technology improvements allowed most of the institutions to utilize large amount of knowledge effectively which is diverse and distributed around the world. But it has become a massive challenge to understand and cover the actual needs of the learners through the existing learning management systems since most of the systems are providing assistance on scheduling and maintenance of the course modules.

Evaluating performance in the e-learning systems becomes a massive challenge because of the different factors which affect the learning models. Many of the qualitative and quantitative factors, which are available in the e-learning framework, highlight different aspects of the students' learning but have not been considered yet for evaluation purposes of the student performances. Therefore a deeper analysis on the behavioural patterns of the students and the factors which affect the student learning in e-learning systems can be used to implement an effective performance model to evaluate the overall performance of each student and it is a much needed requirement at this stage to upgrade the learning capacity in the e-learning education.

Applying data mining methodologies on the educational data has brought a new research discipline where the existing methodologies have been used to model the learning behaviours of the learners. Many of the institutions and other university systems around the globe have tried to overcome the problems of identifying actual student needs through learning analytics. Most of the available system implementations are focused on providing capabilities to the teachers and other knowledge producers to discover the students with difficulties of learning. By analysing their learning environment and other behavioural factors teachers will be able to provide necessary guidance to improve their capabilities or learning capacities.

Other than these early warning capabilities in e-learning systems, many researchers have focused on implementing models to evaluate the overall performance of the students. In many of the researches they based on learning factors to construct an evaluation model to predict the overall performance of the students. Since different factors in the e-learning environment can affect the performance of students in different weights, data mining methodologies was used to model the learning behaviour with high accuracy. But with different learning contexts the same performance evaluation model cannot be used since the data available will not be coped within the implemented model.

Therefore models for evaluating the student performance with acceptable accuracy levels and quality predications still need to be researched more and existing learning analytics should be implemented in such a manner in which they can be used by the knowledge producers with more user friendliness and more interpretation capabilities in an efficient and effective manner.

ACKNOWLEDGMENT

I would like to thank Dr. Shehan Perera for his excellent reviews and guidance given for this literature review which has contributed enormously to produce this paper in such a manner.

REFERENCES

[1] F. Castro, A. Vellido, À. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," *Evolution of teaching and learning paradigms in intelligent environment*, 2007, pp. 183–221.

[2] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, "Data mining algorithms to classify students," *Proceedings of Educational Data Mining*, 2008, pp. 20–21.

[3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Third. Morgan Kaufmann, 2011.

[4] R. Hershkovitz and R. Nachmias, "Learning about online learning processes and students' motivation through Web usage mining," *Interdisciplinary Journal of Knowledge and Learning Objects*, vol. 5, 2009, pp. 197–214.

[5] Eitel J.M. Lauría, "Joshua Baron, Mining Sakai to Measure Student Performance: Opportunities and Challenges in Academic Analytics," 2011.

[6] A. El-Halees, "Mining Students Data to Analyze Learning Behavior: A Case Study," *Department of Computer Science, Islamic University of Gaza PO Box*, vol. 108, 2009.

[7] D. Monk, "Using data mining for e-learning decision making," *The Electronic Journal of e-Learning*, vol. 3, no. 1, 2005, pp. 41–54.

[8] T. Cash, C. Dawicki, and B. Sevick, "Springfield Public Schools Dropout Prevention Program Assessment & Review (PAR)," 2011.

[9] J. L. Hung and K. Zhang, "Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching," MERLOT Journal of Online Learning and Teaching, 2008.

[10] D. García-Saiz and M. Zorrilla, "Comparing classification methods for predicting distance students' performance," 2011.

[11] "Student Individualized Growth Model and Assessment (SIGMA)," 2010.

[12] S. A. Kumar and M. N. Vijayalakshmi, "Efficiency of decision trees in predicting student's academic performance," in First International Conference on Computer Science, Engineering and Applications, CS and IT, 2011, vol. 2, pp. 335–343.

[13] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Efficiency of Machine Learning Techniques in Predicting Students' Performance in Distance Learning Systems," Citeseer, 2002.

[14] T. Hadzilacos, D. Kalles, C. Pierrakeas, and M. Xenos, "On small data sets revealing big differences," Advances in Artificial Intelligence, 2006, pp. 512–515.

[15] M. Ramaswami and R. Bhaskaran, "A CHAID based performance prediction model in educational data mining," arXiv preprint arXiv:1002.1144, 2010.

[16] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting Students' Performance in Distance Learning Using Machine Learning Techniques," Applied Artificial Intelligence, vol. 18, no. 5, 2004, pp. 411–426.

[17] D. Kalles and C. Pierrakeas, "Analyzing student performance in distance learning with genetic algorithms and decision trees," Applied Artificial Intelligence, vol. 20, no. 8, 2006, pp. 655–674.

[18] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Preventing student dropout in distance learning using machine learning techniques," in Knowledge-Based Intelligent Information and Engineering Systems, 2003, pp. 267–274.

[19] B. Minaei-Bidgoli, G. Kortemeyer, and W. F. Punch, "Enhancing online learning performance: an application of data mining methods," in The 7th IASTED International Conference on Computers and Advanced Technology in Education (CATE 2004), 2004, pp 173–178.

[20] B. Minaei-Bidgoli, D. A. Kashy, G. Kortmeyer, and W. F. Punch, "Predicting student performance: an application of data mining methods with an educational web-based system," in Frontiers in Education, 2003. FIE 2003 33rd Annual, 2003, vol. 1, p. T2A–13.

[21] T. Tanner and H. Toivonen, "Predicting and preventing student failure–using the k-nearest neighbour method to predict student performance in an online course environment," International Journal of Learning Technology, vol. 5, no. 4, 2010, pp. 356–377.

[22] Campbell, John P., Peter B. DeBlois, and Diana G. Oblinger, "Academic analytics: A new tool for a new era, " Educause Review 42, no. 4, 2007, pp 40.

[23] Jennifer Natsu, "Advanced Analytics: Helping Educators Approach the Ideal," eSN Special Report, eSchool News, 2010, pp. 17–23.

[24] van Barneveld, Angela, Kimberly E. Arnold, and John P. Campbell, "Analytics in higher education: Establishing a common language," Educause Learning Initiative 1, 2012, pp. 1-11.

[25] E. Galy, C. Downey, and J. Johnson, "The Effect of Using E-Learning Tools in Online and Campus-based Classrooms on Student Performance," Journal of Information Technology Education, vol. 10, 2011, pp. 209–230.

[26] Mamcenko, Jelena, Irma Sileikiene, Jurgita Lieponiene, and Regina Kulvietiene, "Analysis of E-Exam Data Using Data Mining Techniques." in Proceedings of 17th International Conference on Information and Software Technologies (IT 2011), 2011, pp. 215-219.

[27] Bienkowski, M, Feng, M, and Means, B, "Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief, " Office of Educational Technology, U.S. Department of Education, 2012.

[28] C. Beer, K. Clark, and D. Jones, "Indicators of engagement," Curriculum, technology & transformation for an unknown future. Proceedings ASCILITE Sydney, 2010, pp. 75–86.

[29] F.Martin, J.I. Tutty, Y. Su, "Influence of Learning Management Systems Self-Efficacy on ELearning Performance," in i-manager's Journal on School Educational Technology, vol.5, 2009.

[30] Daneshgar, Farhad, Christine Van Toorn, and Daniel Schlagwein. "A Theoretical Model Of E-Learning Ability To Support Attainment Of Students' Graduate Attributes," 2012.

AUTHORS

First Author – Thilina Ranbaduge, Department of Information Technology, Faculty of Information Technology, University of Moratuwa, Srilanka. Email : tranbaduge@uom.lk