A REVIEW ON DATA MINING TECHNIQUES TO PREDICT THE STUDENT PERFORMANCE AND DECISION MAKING IN EDUCATIONAL INSTITUTIONS

KG Abeywickrama[1], WJ Samaraweera, and CP Waduge
Department of Information Technology, Faculty of Computing,
General Sir John Kotelawala Defence University, Ratmalana, SriLanka.
[1]kgabeywickrama@gmail.com

Abstract- Education is significant as it represents the future of a nation. Most of the Sri Lankan educational institutions utilize manual, paper-based systems to manage information which are more time and money consuming. It also reduces the accuracy and work efficiency. Nowadays the commercial world is fast reacting to the growth and potential in data science and as a result, data mining is getting much attention from many researches at present, and data mining assists to discover patterns within enormous amounts of data, stored in databases and data warehouses. Therefore, adapting these techniques will help to find interesting patterns to predict the student performance and to find the grades of students based on their examination results. Through this review paper, an effort is made to investigate a best data mining technique to quantify the student performance to provide benefits for academic staff, administration staff and students. The prediction on performance will provide more precise results and students may receive more accurate predictions which may help to make important decisions in their careers. Most importantly, this will reduce the workload of the administration and will surmount many challenges pertaining to the scholastic field providing a user-friendly environment.

Keywords- Data Mining, EDM, Classification

I. INTRODUCTION

Education system forms the backbone of every nation. Hence, it is important to manage education related processes to provide a strong educational foundation securing the future for everyone. Today educational institutions are not limited to imparting education alone, but also adapting the latest trends in IT to manage and serve the institution resources efficiently to improve the quality of education. At present, numerous studies are taking place in data-mining field. Educational Data Mining is one of the main research fields and also known as EDM. It aims at developing and using algorithms to improve educational results and explain educational strategies for further decision -making. This paper discusses some of the data mining algorithms applied on educational related areas. These algorithms are applied to extract knowledge from educational data and study the attributes that can contribute to maximize the performance.

One of the biggest challenges faced by educational institutions are the exponential growth of educational data and how to apply this data to improve the quality of managerial decisions. Educational Institutions would like to know, for instance, which students will enrol in particular course programs, and which students will need assistance for graduation. Through the analysis and presentation of data collected, on data mining process will help effectively to address the challenges of these students.

Data mining enables organizations to explore and understand hidden patterns in a vast range of databases by using their current reporting capabilities. And these uncovered patterns are then incorporated into data mining models and applied to predict individual behaviour and performance with high precision. In this way, resources

and staff can be allocated by institutions more effectively. Data mining may also, be able to efficiently allocate resources with an accurate estimate of how many students will take necessary actions before he or she drops out.

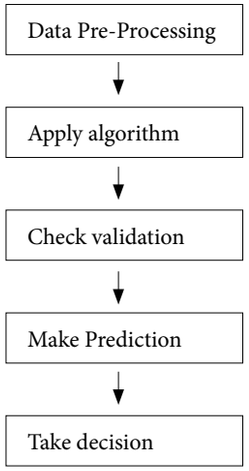

Figure1. Educational data mining Process

Generally, Educational data mining process mainly consists of four stages. During the first phase of the process, the relationships in between data is found by using a training data set. This is done by applying several data mining algorithms such as clustering, classification and association rule mining. In the second phase whatever the relationships identified in the data pre-processing step is checked for validation. As the third step, the validated relationships are used to make predictions. Finally, decisions are taken on the prediction results to filter data.

EDM process facilitates many users, specially administrators, educators and researchers. All the stakeholders will be benefited by data mining as it will ease their work in decision - making, do changes in teaching processes, allocating resources for institutions. So, the educational processes in institutions may provide best results.

II. LITERATURE REVIEW

Educational field has realized the importance of using data mining especially when examining students 'learning performances. Data mining can be defined as discovery of knowledge, which involves searching for unexpected but interesting patterns within large amounts of data, normally stored in databases and data warehouses.

Recently, data mining focused on analysing educational related data to develop models to improve student's learning experiences and enhance institutional efficiency. Therefore, data mining help educational institutions to provide high-quality education for their students. Applying data mining in education also known as educational data mining (EDM), which enables to better understand how students are learning, and determine how it improves educational outcomes. (Maqsood Ali,2013)

Student related information management, in an educational institute becomes tiresome each year, so it's better to have an automated system for managing such information. (Dalgade, et al., 2016)

Authors (NWOKE and IGBOJI, 2015) have developed a powerful offline and online school management software which plays vital roles in gaining competitive advantages. In addition, they explained that with proper planning and management, student records can be a valuable resource for many people, ranging from parents and local school officials to researchers and policymakers.

Educational data mining is emerging as a research area with a collection of computational and psychological methods and research approaches to understand how students learn. Therefore, if the academic planners can organize a conference for final year students then it will improve the overall productivity. (Dwivedi and Singh, 2016) provide an overview of the survey of student performance which helps academic planners in an institute to guide students to choose their right carrier in which their talents and commands are good. They have suggested that it's better to prefer algorithm like decision tree induction, association, logistic regression and naïve Bayes to filter educational data for placement predictions. (Dwivedi and Singh,2016)

(Bhardwaj and Pa, 2011) conducted a notable research on data mining using naive Bayes classification algorithm to predict the performance of students with 13 classification variables. According to his research observation, Data mining techniques are applied for vast amount of data to discover hidden patterns and relationships helpful in decision - making. Bayesian classification is one of the most useful method of data mining used on student databases to predict the students' performance on the previous year database. It helps students and the teachers to improve the class of the student and to identify students who need special attention to reduce the failing rate and to take appropriate actions at the right time. Moreover, that study shows university students' performance do not always depend on their own effort but, also depend on social, psychological and other environmental factors.

Evaluation is an important element in teaching and learning. The arrival of the Internet and related technologies have made online assessment systems feasible and popular in education and training. Evaluations can be formative or summarized. Computer based interactive multiple-choice question exams are proposed for formative evaluations while essay type exams conducted for summarize assessments. (Gogri et al., n. d) proposes a system of performance assessments for students. And the data generated from these evaluations are used for data mining. Classification technology C4.5 is used in decision tree showed that formative assessment, leading to better development of students while summarize assessment compels students to focus on how many marks have they secured rather understanding the content they are studying. Therefore, at the end they are suggesting that formative assessment strategies are more effective in improving quality of students learning and leading to better development of students.

There are various open source tools which are specialized for data mining. Some of them are Weka, RapidMiner, Orange, Knime, DataMelt, etc.(Kabakchieva, 2013) carried a research project using the WEKA software which is based on the C4.5 decision tree algorithm. As it is freely available to the public and is widely used for research purposes in data mining field. This research project reveals the high potential of data mining applications for university management by analyzing the performance of different data mining algorithms. And the results achieved from the research revealed that the decision tree classifier (J48) performs best followed by the rule learner (JRip) and the kNN classifier. The Bayes classifiers are less accurate than the others.

(Nithya et al., n.d.) has done a survey on educational data mining. They have discussed every algorithm which used in education mining. And have found that these algorithms show a remarkable improvement in strategies like course outline formation, teacher student understanding and high output and turn out a ratio.

(Lan and Li, n.d.) in 2011 published a research paper to improve the association rule mining algorithm. This algorithm mines the association rules of the courses, identifies the relationship of students, selected courses and provide basis for planning and curriculum classification. At the end of the experiment they have concluded that the association rule mining algorithm not only achieve the function of data mining, but also increases the mining speed to a certain extent and its simple and easy to implement.

(Baepler and Murdoch, 2010)stated that the new directions of academic analytics and data mining will produce new opportunities for collecting, analysing and reporting student's data. And those data can be used to redesign the course content, assessments etc. After series of experiments they have concluded that data mining is truly effective, influence curricular advancement and provide instructional choices for both students and the faculty.

(Yu et al., 2010) shows that data mining techniques can be used to study the factors that influence the retention of a university student. They have explored this using three data mining techniques namely, classification trees, multivariate adaptive regression splines(MARS) and neural networks. The data mining tools used in this study are used to provide some vision into various aspects of student retention that were not revealed before and encouraged researchers to investigate further.

(S. Abu-Oda and M. El-Halees, 2015) applied different data mining techniques with the purpose of examining and predicting students' dropouts through the university programs. They have used Decision Tree(DT) and Naïve Bayes(NB) classifiers and found out that accuracy of NB is higher than DT. This study also discovered hidden relationships between student dropout status and enrollment persistence by mining some frequent cases using F-P growth algorithm. Finally, they have concluded that mastering algorithm analysis courses a great effect on predicting student persistence in the major and decrease student likelihood of dropout.

(Agarwal et al., 2012) has taken student data from a community college and different classification approaches have been performed and a comparative analysis has been done. They have established that Support Vector Machines(SVM) as the best classifier and Radial Basis Kernel as the best choice for SVM. They had analyzed the

data available on a student's academic record and student likelihood in terms of placement may be predicted based on the admission test results, quantitative ability marks and verbal ability marks by using decision tree approach and adopting decision rule approach. And finally concluded that Data Mining could be used to improve the process of business intelligence including the education system in order to enhance the efficacy and overall efficiency by optimally utilizing the resources available.

III. METHODOLOGY

Education in any institution can be upheld by applying data mining technology, because Data mining helps to improve the standards and efficiency of the educational systems.

Various algorithms and techniques are used for knowledge discovery from databases. Classification, Clustering, Association Rule Mining and Sequence Analysis are some of the models used to implement a model.

Hence, identifying research questions will help to identify the scope of the study and to carry out a systematic review to support the researches.

A. Research Questions

Research questions are important to understand the study of predicting student's performance.

Two most important research questions are suggested in this study are:

- Q1: What are the important parameters, or attributes that are used to predict the student's performance?
- Q2: What type of algorithms are used to predict student's performance?

Next section of the study will discuss the proposed research questions that will be useful to predict students' performance.

B. The important parameters used in Student's Performance Prediction

When predicting student performance, the main predicated parameter will be their examination marks or the average score.

Before applying an algorithm, data has pre-processing step. If we assume, score variable have five distinct categories as grade" A"," B","C"," S" and" F". Students those who have scored in the range of 85 and 100 are belong to grade A, scores range between 75 to 84 belong to grade B, grade C in the range between 65 to 74, while S in the range between 55 to 64 and F in the range below 55.

To build a model that would classify the students into the five classes depends on the data collected and different algorithms are applied for predictions such as decision trees, Bayesian networks etc. We can train different models to predict student performances. In this way we can test the accuracy of each model.

C The prediction algorithms used for student performance

1) Naive Bayes Classifier

Bayes Classification is used to estimate the probability of a certain property in the data set. Naïve Bayes Algorithm is descriptive and a predictive type of algorithm. And it is easy to use from other approaches as, it only requires a small amount of training data to evaluation, because only one scan of the training data is required.

The Naïve Bayes algorithm classifies the cases by considering the independent effect of each attribute to the classification. And the ultimate precision is determined by the results achieved.

2) Decision Tree Classifier

Decision tree is a tree like structure that is used to describe the relationship between properties and their significance.

Decision tree algorithm is easy to understand and interpret. And works well for both quantitative and complex categorical data.C4.5 can be used with decision tree algorithm to obtain information which requires more attention.

3) Rule Learners Classifier

There are two algorithms, OneR and JRip. OneR generates one-level decision tree which is a simple and cheap method with high precision for characterizing the structure in data. And JRip implements the RIPPER algorithm where classes are studied in increasing size and generate set of rules for the class using incremental reduced-error pruning.

D. Logistic Regression

Logistic regression is used to describe data and to predict the dependent variables that explains the relationship between one dependant binary variable and one or more nominal variables ordinal, interval or ratio-level independent variables.

Table 1. Performance comparison between the classifiers

Author	Technique	Result
Oktarani Nuruk Pratiwi 2013	OneR J48 Naïve Bayes	78.66% 79.61% 76.75%
Ajay Shiv Sharma,S.S 2014	Logistic Regression	83.33%
Vikas Chirumamilla, B.S 2014	C4.5	77.78%

IV. DISCUSSION

In the field of education, for predicting the student performance, we can use data mining technology. It helps to make decisions on exam results and many more related to students. And lecturers can guide the students to the correct path to improve their career after analysing results.

From the review carried on educational data mining showed that the accuracy of Logistic Regression is higher

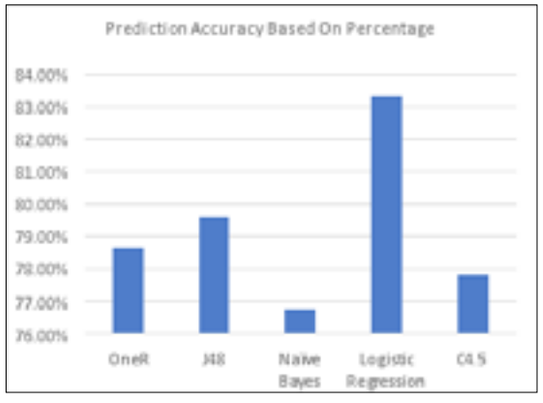

Figure 2. Distribution of Algorithms: Performance of the algorithms

than all the other algorithms. Therefore, it's better to use Regression and decision tree classifier as it performs best; followed by the rule learner (JRip) classifier to predict the student performance.

V. CONCLUSION

This paper analyses different data mining techniques to predict the student performance and find the grade of students. When predicting student performance, the examination marks or the average score have been used as a main data set by many researches. According to the research review, accuracy of logistic regression is higher than all the other algorithms for predicting student performance. Therefore, it's better to use logistic regression to predict the student performance. Further, this study will help students to make important decisions in their career path and lecturers will be able to identify weak students who need special attention to reduce the failing rate. Therefore, if all the scholastic centres can use data mining techniques in their resource management systems, it will improve the overall productivity and will enhance the future of students, improving learning and teaching processes. In Conclusion, I hope this review paper inspires the researchers to explore further on educational data mining and help to measure the student performance in a systematic way.

REFERENCES

Agarwal, S., Pandey, G.N., Tiwari, M.D., 2012. Data mining in education: data classification and decision tree approach. Int. J. E-Educ. E-Bus. E-Manag. E-Learn. 2, 140.

Baepler, P., Murdoch, C., 2010. Academic Analytics and Data Mining in Higher Education. Int. J. Scholarsh. Teach. Learn. 4. https://doi.org/10.20429/ijsotl.2010.040217

Bhardwaj, B.K., Pa, S., 2011. Data Mining: A prediction for performance improvement using classification. IJCSIS Int. J. Comput. Sci. Inf. Se Curity 9.

Dalgade, P.D.M., Panday, A., Negi, G., Popat, N., 2016. A RESEARCH PAPER ON STUDENT INFORMATION AND SCORE MANAGEMENT SYSTEM (SISMS) [WWW Document]. URL http://troindia.in/journal/IJACCCS/vol2iss2/13-16.pdf (accessed 10.28.17).

Dwivedi, T., Singh, D., 2016. Analyzing Educational Data through EDM Process: A Survey. Int. J. Comput. Appl. 136, 0975–8887.

Gogri, M.H., Shaikh, S.A., Iyengar, V.V., n.d. Evaluation of Students Performance based on Formative Assessment using Data Mining - pxc3886623.pdf [WWW Document]. URL http://research.ijcaonline.org/volume67/number2/pxc3886623.pdf (accessed 9.21.17).

Kabakchieva, D., 2013. Predicting Student Performance by Using Data Mining Methods for Classification. Cybern. Inf. Technol. 13. https://doi.org/10.2478/cait-2013-0006

Lan, A., Li, J., n.d. College information system research based on data mining [WWW Document]. URL http://www.icmlc.org/icmlc2009/080_icmlc2009.pdf (accessed 10.28.17).

Nithya, D.P., Umamaheswari, B., Umadevi, A., n.d. A Survey on Educational Data Mining in Field of Education.

NWOKE, B.O., IGBOJI, K.O., 2015. Automated School Management System – Recipe for Viable Educational System in Developing Countries. Int. J. Eng. Trends Technol. IJETT 25.

S. Abu-Oda, G., M. El-Halees, A., 2015. Data Mining in Higher Education : University Student Dropout Case Study. Int. J. Data Min. Knowl. Manag. Process 5, 15–27. https://doi.org/10.5121/ijdkp.2015.5102

Yu, C.H., DiGangi, S., Jannasch-Pennell, A., Kaprolet, C., 2010. A data mining approach for identifying predictors of student retention from sophomore to junior year. J. Data Sci. 8, 307–325.