RESEARCH ARTICLE OPEN ACCESS

A Prediction Model of Students level at Graduation Using Educational Data Mining

Haitham Alagib Alsuddig [1], Piet Kommers [2]
College of Computer Science and Information Technology [1],
Sudan University of Science and Technology, Sudan
Faculty of Behavioural Sciences [2], University of Twente, Netherlands

ABSTRACT
This paper aims to control the quality of the educational process inputs and improve its mechanisms to ensure the outputs of the representative at the success and fail of the student. Sudan University of Science and Technology College of Computer Science was selected to be the field of application. The paper presented a set of indicators that help the university administration to diagnose the effectiveness of some elements of the inputs of the educational process and guide them towards laying the foundations for the development of these elements, including increasing effectiveness. This was done through the use of exploration methods in educational data and the concept of cloudy logic and mathematical logic to provide different models are applied according to the type of issue being addressed,Different methods and techniques of data mining were compared during the prediction of the Students level at Graduation, applying the data collected from the surveys conducted from the SUST the Faculty of Computer Science year 2012-2013,2013-2014,2014-2015among first year students and the data taken during the enrollment. The success was evaluated with the passing grade at the exam. The achieved results from high school and from the result of the first year have an effect on student level, were all investigated. In future investigations, with identifying and evaulating variables associated with process of studying, and with the sample increase, it would be possible to produce a model which would stand as afoundation for the development of decision support system in higher education.

Keywords: - Data Mining, Classification, Prediction, Student Success, Higher Education

I. INTRODUCTION

Education is the foundation of nation building and development. Accordingly, the last decade has seen increasing interest in finding ways to improve the quality of teaching and to improve teaching and learning. The development of educational information systems and the proliferation of learning techniques led to the availability of many data on the educational process, which led researchers to think the need to apply methods of data mining to extract useful information from the data provided by the educational systems, and led to the emergence of independent research areas such as exploration in educational data. Education Data Mining (EDM) (Learning Analytic) Learning data mining techniques seek to access educational data repositories and extract useful information that helps to better understand the educational process and improve the teaching and learning process. In the educational data the same approach used in traditional methods of data mining of the need to understand the environment to be dealt with and then collect the data and then cleaned and arranged and select the techniques that can be applied and finally interpret the results and verify the validity of the techniques applied.

The aim of this study is to propose a model to predict students expected to graduate at a critical level, based on data available to students after the end of the first academic year so that the Department of Education can follow and correct from an early date. In this model we will answer the following question: Can we find a model that will help the Department of Education to identify students who are expected to perform poorly when graduating to monitor and correct from an early date? This issue is part of the issues that predict the level of students upon graduation based on a set of indicators.

Experiments have shown that graduates with low rates find it difficult to recruit and find good university admissions to complete their studies. Thus, there is a need to limit students who are expected to graduate at a critical level to help them correct early. By securing an appropriate academic plan after discovering flaws and weaknesses, the early detection was the greater the chance of correcting the path.

Prediction of student's graduation level at an average starting from the early stages of his or her academic career helps the educational administration to develop appropriate methodologies to improve the level of students expected to perform poorly or to support students who are expected to perform well.

II. RELATED WORK

One of the studies, presented a comparative study on the effectiveness of educational data mining techniques to early predict students likely to fail in introductory programming courses. Although several works have analysed these techniques to identify students' academic failures.

The study evaluated the effectiveness of four prediction techniques on two different and independent data sources on introductory programming courses available from a Brazilian Public University: one comes from distance education and

the other from on-campus. The results showed that the techniques analysed in this study are able to early identify students likely to fail, the effectiveness of some of these techniques is improved after applying the data pre-processing and/or algorithms fine-tuning, and the support vector machine technique outperforms the other ones in a statistically significant way (Costa, 2017).

There is another study had used classified students by using genetic algorithms and regression method to predict their final grade (Minaei-Bidgoli, 2003).

Another study used Several data mining techniques like decision tree, Neural network, Linear Discriminant Analysis. 20% variables showed significant correlations with academic success and 80% rate of correct classification in predicting the success or failure of students. Students' performance were categorized in five groups: ―Very good‖, with a high probability of succeeding; ―Good‖ students, who are above average and with a little more effort can succeed with good grades ; ―Satisfactory‖ students, who may succeed; ―Below Satisfactory‖ students, who require more efforts to succeed; and ―Fail , who have a high probability of dropping out (Dekker, 2009).

In other research, Saurabh build a model using data mining methodologies to predict which students would likely drop out during their first year in a university program. That study used the Nave Bayes classification algorithm to build the prediction model based on the current data. The result of the system was promising for identifying students who needed special attention to reducing the dropout rate (Pal, 2013).

One of the studies focused on predicts student's characteristics or academic performances in various educational institutions. This paper focus on students' performance as a slow learner or fast learner. For that they applied various data mining techniques and compare the accuracy based on students' attributes [6]. For assessing the goodness of a predictor, an extensive study on the student data set was conducted by applying five individual classifiers J48 (J48), Bayesian Net (BN), Neural Net (NN), Decision Tree (DT), and Naïve Bayes (NB) (Asif, 2017).

Many studies around the world have been interested in applying data mining algorithms to discover knowledge in universities ,One of the most important of these studies is a study concerned with the applications of data mining in the field of higher education, Focused on the input of the educational process and its outputs and how they affect each other, The study used the method of neural networks to explore data, The results showed diverse relationships between curricula , multiple hours , the nature of students and between the graduates and the jobs they occupy, As well as other useful conclusions for decision-makers at universities

Another study was conducted in Ethiopia In Debre_Markos University study has shown that data mining techniques can be applied by higher education institutions or universities in determining student failure/success rate so that managing students' enrolment at the beginning of the year, assist students before they reached risk of failure, effective resource utilization and cost minimization, helping and guiding administrative officers to be successful in management and decision making. The study applied data mining technology to the data of university students for the purpose of forecasting The success or failure of students, The study used CRISP methodology The analysis was carried out by the WEKA program and the forecast model was built The study found the main vgclass, number of courses given in a semester, and field of study are the major factors affecting the student performances (Asif, 2017).

Also, there are studies discussed that Student performance in university courses is of great concern to the higher education managements where several factors may affect the performance. This study is an attempt to use the data mining processes, particularly classification, to help in enhancing the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses (Gulati, 2012).

Ramaswamy N has written research paper which focused on predicts student's characteristics or academic performances in various educational institutions. This paper focus on students' performance as a slow learner or fast Lerner. For that they applied various data mining techniques and compare the accuracy based on students' attributes. For assessing the goodness of a predictor, an extensive study on the student data set was conducted by applying five individual classifiers J48 (J48), Bayesian Net (BN), Neural Net (NN), Decision Tree (DT), and Naïve Bayes (NB) (Ramaswami, 2014).

Another research used data mining techniques for predicting the students' graduation performance in final year at university using only pre-university marks and examination marks of early years at university, no socio-economic or demographic features are use. The result of the study shows that can predict the graduation performance in a four-years university program using only pre-university marks and marks of first- and second-year courses, no socio-economic or demographic features, with a reasonable accuracy, and that the model established for one cohort generalizes to the following cohort. It makes the implementation of a performance support system in a university simpler because from an administrative point of view, it is easier to gather marks of students than their socio-economic data. The result also shows that decision trees can be used to identify the courses that act as indicator of low performance. By identifying these courses can give warning to students earlier in the degree program (Olalekan, 2017).

The study propose an approach to predict the performance using data mining techniques, the study also shows that the most important attributes which most affected the performance of students who consume the alcohol during their study are the previous grades which is gained by students and other attributes are absence in the class, father's

job, mother's job, extra educational support, extra paid classes within the course subject, wants to take higher education, reason to choose this institution and also some other attributes (Pal S. a., 2017).

III. METHODOLOGY

This paper finding that there are several different methods of data mining, and the choice of the appropriate method depends on the nature of the data under study and its size like analysis Correlation, decision tree, genetic algorithms, virtual theory networks, raw group path, neural network, statistical analysis and There are several tools to explore the data, the most important of which are Summarization, Classification, Prediction, Clustering, Rule Analysis, and Change and deviation detection.

Through the study I found that the most appropriate tools used in the educational data mining is the prediction tool, some of the traditional tools used in Predictions are, for example, regression and differential analysis. The new methods include correlation rules, decision tree, neural networks, and genetic algorithms.

After looking in previous studies, I found that the best methodology for applying the Education Data Mining and predictive model as show it in the figure below:

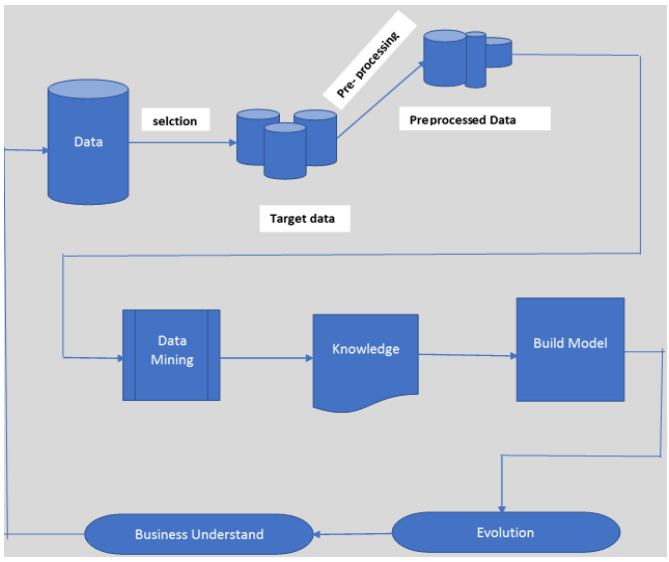

Fig. 1. The steps of Methodology.

The figure above illustrates the stages and steps of Data Mining and modelling

I have collected data from the Directorate General of Admission - Ministry of Higher Education and Scientific Research and Faculty of Computer Science, Sudan University of Science and Technology for three batch 2012, 2013 and 2014. The total sample 347 students, after that I had Built dataset by using of Microsoft Access database program such as join and query tools

In pre-processing phase, noise data that is insignificant is removed, and inconsistent data and inconsistent data are deleted.

At Data Mining stage I used correlation rules and take the student's mark in high school, First Year and the student's mark in mathematics as Inputs for the prediction model.

. Then I have been suggested a predictive model for students who are expected to graduate at a critical level by using an adaptive neural fuzzy inference system.

IV. RESULTS

By applying the Pearson correlation coefficient of the candidate variables for the model I had selected the inpupt of the models, I found that there is a correlation to all the selected income elements especially in the student's mark in high school, First year and math. That mean can take the student's mark in high school, First Year and the student's mark in mathematics as Inputs for the prediction model.

The next step it was Building the Predictive Model, in this model we will rely mainly on inputs representing the intermediate students' marks. These inputs will be relied upon to predict the student level of a continuous quantitative nature that is best converted into a fuzzy model. The adaptable fuzzy nervous system has been selected as a key tool for predicting the student's final rate. Since we are looking for students who are expected to graduate at a critical level, we will propose a model that begins by identifying the appropriate inputs and applying the fuzzy nervous system in order to predict the final rates of the students and then applying the dummy

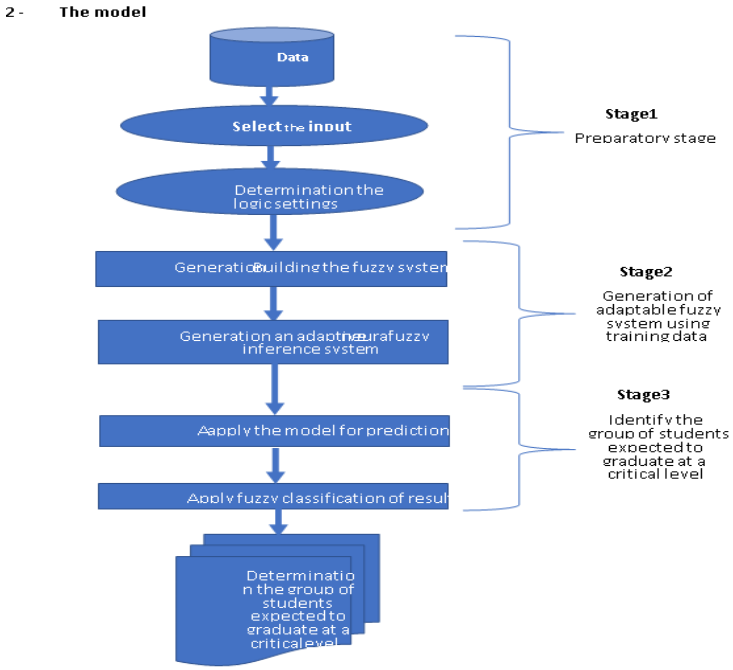

Fig. 2. The Proposed predictive model.

Stage 1: Preparatory

1-Identification of Inputs: The input of the model is determined by proposing a set of candidate inputs and then selecting the most significant in determining the level of the student after graduation, depending on the results of the indicative methods and the application of statistical indicators such as the correlation coefficient value.

2 - Determine the most appropriate settings for building fuzzy system: by determining the number of fuzzy sets for each income and then determining the form of affiliation and averages for each cloud group, depending on the experience and methods of evidence and the expert can change the settings as appropriate.

3 - Determine the mechanism to classify students according to their level: where the expert identified a number of cloudy groups to classify students to their levels according to their expected rate with the identification of the appropriate belonging to these groups.

Stage 2: Building the adaptive fuzzy nervous system using training data

1. Building the fuzzy sets system: according to what was specified from the first phase.

2.Generation of adaptable fog system using training data.

Stage 3: Determine the group of students expected to graduate at a critical level.

1. Preparation of inputs and application of the system for the purpose of prediction: We will use the fuzzy nervous system that was built in the previous stage for

The purpose of predicting after the processing of inputs, where the output will be a numerical series representing the students' expectations.

Apply the classified classification on the results: After the conversion of the expected Rates to the cloudy form, the data is classified in accordance with the classification of the cloudy classification to be identified in the first stage.

Determining the group of students expected to graduate at a critical level: by selecting students who belong to the level corresponding to graduation at a critical level according to the classification criteria.

In testing phase of the model, I used Group Method of Data Handling application (GMDH) that uses adaptive neural network technology, which I flowed to build the model and I had got results:

TABLE 1

The sample of results after Appling the Predictive model.

Student ID	Math	Secondry dgree	First year	Predictions	Evaluation
1	86	84.3	2.93	3.03	0
2	89	84.9	2.6	2.84	0
3	78	79.4	2.62	2.50	0
4	85	84.9	2.91	3.03	0
5	85	82.7	2.55	2.68	0
6	85	83.4	2.42	2.32	critical level
7	71	86	2.48	2.42	0
8	83	71.7	2.9	3.03	0
9	84	83.1	2.82	3.03	0
10	86	83	2.7	2.84	0
11	79	84.1	2.55	2.50	0
12	86	79.9	2.72	2.84	0

The figure above shows the some of the result after apply the model, for example, through the results, there are 45 students expected to graduate at a critical level.

V. DISCUSSION

Through the application of the rules of correlation technique, the most influential factors affecting the students' failures are the subject of programming methods.

Through the application of adaptive neural network technology, a model was obtained to predict the final results of the students. Graduation assists decision makers from taking precautionary measures to prevent their occurrence. The result of the first year was the first and most important factor in determining the forecast process

Through the application of grouping technology, the group of students expected to graduate has been identified at a critical level, which helps the teacher lead and guide them from early on.

In the future we aspire to expand the research circle to include all faculties of the Sudan University of Science and Technology, so as to see the full picture and help the Ministry of Higher Education to make strategic decisions to improve the academic reality in Sudan.

VI. CONCLUSION

Data mining systems use mathematical, statistical and intelligent methods for building future forecasts and exploring behavior and trends, allowing for estimation of decisions Correct and take them in time. The core of decision support systems is data mining and prediction, early warning and ddecision support systems synthesize available data with personal visions of the decision maker, is done.

There are seven types of data mining, namely: analysis

Correlation, decision tree, genetic algorithms, virtual theory networks, raw group path, neural network, statistical analysis.

The prediction is similar to classification or estimation, except that data are classified as predicting future behavior or

estimating its future value. Where the predicted dependent variable is a quantitative variable. Some of the traditional tools used in forecasting are, for example, regression and differential analysis. The new methods include correlation rules, decision tree, neural networks, and genetic algorithms.

REFERENCE

[1] Asif, R. a. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers \& Education, 113*, 177--194.

[2] Costa, E. B. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247--256.

[3] Dekker, G. W. (2009). Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining*.

[4] Gulati, P. a. (2012). Educational data mining for improving educational quality. *Int. J. Comput. Sci. Inf. Technol. Secur, 2*, 648--650.

[5] Minaei-Bidgoli, B. a. (2003). *Predicting student performance: an application of data mining methods with an educational web-based system* (Vol. 1). 33rd Annual Frontiers in Education, 2003. FIE 2003.: IEEE.

[6] Olalekan, A. M. (2017). Prediction of Graduating Students for Tertiary Institutions Using Data Mining Technique.

[7] Pal, A. K. (2013). Classification model of prediction for placement of students. *International Journal of Modern Education and Computer Science*, 49.

[8] Pal, S. a. (2017). Performance Analysis of Students Consuming Alcohol Using Data Mining Techniques. *International Journal of Advance Research in Science and Engineering, 6*, 238--250.

[9] Ramaswami, M. (2014). Validating predictive performance of classifier models for multiclass problem in educational data mining. *International Journal of Computer Science Issues (IJCSI), 11*, 86.