A Survey on various Issues & Challenges on Student's Performance Dataset

Reena Yadav[1] Mrs. Rajni Kori[2] Dr. Shiv K. Sahu[3]
[1]PG Student [2]Assistant Professor [3]Head of Dept.
[1,2,3]Department of Computer Science & Engineering
[1,2,3]LNCTE, Bhopal (M.P.) India

Abstract— Student attrition is a problem most higher education institutions have to arrangement with as there are both financial and human costs associated with it. Different researchers have studied it primarily using theoretical models and then conventional models. In recent times research on student attrition using data mining became extensive and in most cases attempts to predict the phenomena almost always using factors within the students. Here in this paper a survey of all the existing techniques that is implemented for the prediction of Student's Performance.

Key words: Educational Data Mining, Data Mining (DM), Knowledge Discovery in Databases (KDD), Classification, Learning Analytics (LA)

I. INTRODUCTION

As with other organizations, the level of support obtained by the use of information technology (IT) in higher education institutions has been diversified. It can be very basic only providing support for small repetitive tasks such as word processors for certificates emission or spread-sheets for calculating final classifications. On the other hand, it can be process oriented, by the form of isolated applications to support specific processes such as student's enrolment. More recently, IT support has been fully integrated in the form of systems that support most of the processes of the institution and by using a common repository of information. In recent years, three emerging fields are using data and technology approaches to improve Education and Learning. Academic Analytics, which uses a business intelligence approach to Education in order to improve decision making and organizational efficiency; Learning Analytics, which looks to empower the actors of the learning process; and Educational Data Mining, which is a branch of Data Mining specialized on Educational needs from the learner or the organization. In educational settings, Data Mining techniques have been applied in both, Learning and Administrative/ policy-oriented issues [1, 2]. In Learning, the process can be split into learner-oriented and educator-oriented. In the first one, the focus is on supporting the student to learn more effectively by suggesting new contents; in the latter, the goal is to provide the educator a tool to empower him so he can guide the learner more effectively. Classification techniques are used [3] for prediction of student performance in distributed environment. Data mining methods are often implemented at many advance universities today for analyzing available data and extracting information and knowledge to support decision-making.

Data Mining (DM) or Knowledge Discovery in Databases (KDD) is the automatic extraction of implicit and interesting patterns from large data collections. It has been used and applied in a diversity of grounds such as industries, government, military, retail and banking. Recently it has received much attention in educational contexts [4]. The first efforts were carried out by Campbell and Oblinger [5] who introduced "academic analytics" defining it as the use of statistical techniques and data mining methods aiming to help faculty and advisor's in becoming more proactive in the identification of students possibly at risk and reacting accordingly. This kind of research pointed the path to the acknowledgement of a new sub-topic of data mining called Educational Data Mining (EDM). Baker and Yacef [1] provided a formal definition as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. It is worth noting that this definition does not clearly mention the expression "data mining", however, it does so implicitly, as it emphasis the exploratory nature of the process, which is one of its key characteristics. Therefore, educational data mining was chosen over academic analytics because it encompasses nearly all types of data in educational institutions and focuses on applying data mining tools and techniques [6] Moreover, it highlights the importance of analysing educational data for the development of models for improving learning experiences and improving institutional effectiveness. As different data sets have been used and recently research relies exclusively on academic data as source of data. Data sets can relate to a whole institution or just focus on a specific programme or group of courses. Comparisons are not easy to perform because the data set is from within the institution where research is carried out, therefore showing the noticeable lack of a baseline data set for comparing results within different researchers.

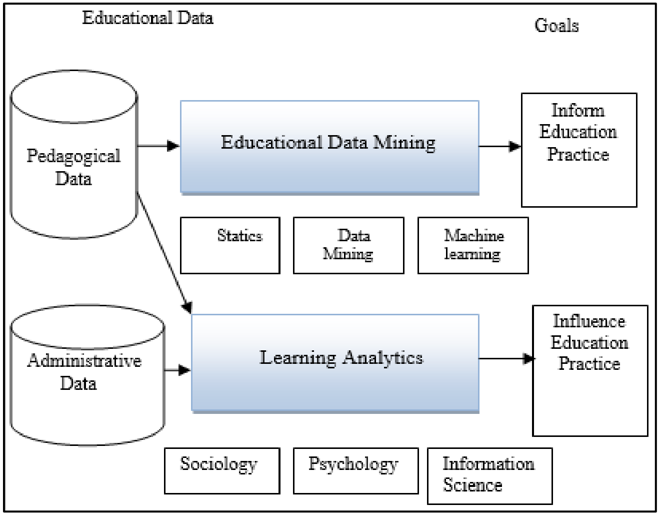

Fig. 1: Big data in education

In Bienkowski's [7] view, educational data mining can also be considered a sub area of big data in education, the other one being learning analytics. The two fields have had different origins and settled as distinct research areas.

Educational data mining develops methods and applies techniques from statistics, machine learning and data mining to analyse data collected from administrative services. From these it should be possible to test learning theories, develop algorithms and models to inform education practice. Learning analytics besides the former, also applies techniques from information science, sociology and psychology to analyse both pedagogical (teaching and learning environment) and administrative data (academic record facts), with a goal to influence education practise. These two lanes of research are depicted in Figure 1. While making sense in theory, in hands-on terms, this distinction seems rather blurred as the goals of both fields are very intertwined. As a consequence, educational data mining ended up being used overall.

II. Theoretical Background

Technology has been an enabler for education. The first thoughts of this influence might be commonly related to a way for communicating, for delivering content or interacting with students by using video and other media to support a message, or creating virtual learning environments that facilitate communication; there is also the possibility to maximize access to education with online courses. However, these are not the only possibilities, Education, as many other fields, can also be improved by the use of data and analytics to enable a better decision making.

Analytics involves the use of data and quantitative analysis in the decision making process. This is supported by the recent increase of volumes of data and computational resources, which is changing the paradigm of science, from theoretical models, to computational models and finally to a data-intensive science [8]. New tools coming from Data Mining, Machine Learning or Statistics can be applied during the process of exploratory analysis by discovering new patterns that possibly were not considered by experts, and reducing the number of traditional data collection - hypothesis testing techniques to only a few interesting patterns[9].

III. Data Mining Techniques in Educational Systems

Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so that they may be used in an enterprise's decision making process. It is usually used by business intelligence organizations, and financial analysts, but it is increasingly used in the sciences to extract information from the enormous data sets generated. Bringing together the categories from both taxonomies, data mining techniques can be listed as follows:

– Prediction, classification and regression - Prediction typically involves creating a model to find the value of a variable whose value is not known as a result of the combination of values of other variables (whose values are known). Prediction has been widely used to understand the causes of student attrition and behaviours of students in e-leaning systems, although presenting challenges in interpretation. Predictions are carried out through classification and regression algorithms. In classification problems, the classes are predetermined [10] and are limited to a finite set, previously made by a human. A training set of data is hand-labelled with these classifications. The algorithm task is to induce a mathematical model from the available features in the training set of data in order to show a relationship to the classification given. The model is then tested in a different set of data. Its predictive performance is evaluated by how close to the classification previously hand-labelled the automatically given one is. There are several algorithms which differ in the approach used in analysing data. Examples include Naïve Bayes, k-nearest neighbour (k-NN), neural networks and Support vector machines (SVM). In regression problems the goal is to predict the value of a continuous numeric variable.
– Outlier detection - An outlier is an instance which appears to be inconsistent with the remainder of the set of data which belongs to [11]. Outlier detection aims to find these instances and the causes to it. They may be due to errors in loading the data into a system, errors in measurement of data, data belonging to another dataset or even data measured correctly but related to a rare event [12] In educational data mining, outlier detection has been used to find students with unusual behaviours, such as having learning problems or potentially gifted students.
– Clustering - Clustering are a set of techniques for finding groups of related instances in a data set by analysing the similarity in characteristics (translated into attribute values). And is usually used to mean segmentation. Using clustering, students can be grouped based on educational background, age, areas of interest, specialization and so on [4]. There are several clustering algorithms. Some can either start with no prior hypotheses about clusters in the data, such as the k-means algorithm [13] with randomized restart, or start from a specific hypothesis, possibly generated in prior research with a different data set. [14] Also a clustering algorithm can determine that each data instance must belong to exactly one cluster or that some instances may belong to more than one cluster or even no clusters at all.
– Relationship mining - Involves the discovery of relationships between variables in large data sets and encodes them as rules for reuse in other contexts. While typical examples of applications are recommender systems by market baskets analysis, is has been used to investigate in learning management system such questions as why students' use of practice tests decreases over a semester of study [7] Another form of relationship mining is sequential pattern mining. In this techniques, the goal is to find rules that capture the linkages between the occurrences of sequential events such as temporal sequences of errors made by student's in an online system that are followed by attempts to seek help. It should be pointed out that results of relationship mining are only useful when they reveal unexpected rules on the data.

IV. Literature Survey

In this paper [3] classification methods are used for prediction of student performance in distributed situation. Data mining methods are frequently applied at many progress universities at present for analyzing available information and extracting information and understanding to support decision- making.¬

while it is important to have models at local stage their effects makes it complicated to remove knowledge that can be valuable at the global height. Consequently, to support decision making at this region it is significant to simplify the information contained in those models; precise classifier technique can be utilized to simplify these rules for global model. Predicting student performance is valuable to produce well-organized and high-quality feature student work force ,by predicting student at possibility and give them enhanced training to get better their performance will definitely valuable for their individual results and also for academic institution outline.

Delen (2010) [15] research using several data mining techniques i.e. both individuals as well as ensembles, developed analytical models to predict and explain the reasons behind freshmen student attrition. The data mining algorithms used were decision trees, SVM, neural networks and logistic regression. Institutional data of about 16000 freshmen students, from 2004 to 2008, provided the data set (from a single institution, a comprehensive public university located in the mid-west region of the USA). As with previous research, results confirm that ensembles of classifiers performed better than individual models using bagging, busting and information fusion. The sensitivity analysis of developed models revealed that the educational and financial variables are among the most important predictors of the phenomenon. Among the four individual prediction models used in this study, SVM performed the best, followed by decision trees, neural networks and logistic regression. As with previous research, it was recognized that decision trees are the best algorithms to provide a justification for a specific predicted outcome.

Kabakchieva (2012) [16] wanted to reveal the high potential of data mining applications for university management in order to contribute to more efficient university enrolment campaigns and attracting the most desirable students. The research was focused on the development of data mining models for predicting student performance, based on their personal, pre-university and university-performance characteristics. The algorithms tried were OneR rule learner, a decision tree, neural network and k-nearest neighbour (k-NN) classifiers. The dataset used for the research includes data about students admitted to the University of National and World Economy of Sofia (Bulgaria) in three consecutive years. The main conclusions were that the highest accuracy is achieved for the neural network model, followed by the decision tree model and the k-NN model. The data attributes related to the students' university admission score and number of failures at the 1st year university exams are among the factors influencing most the classification process.

Bogard et al. (2012) [17] research had the purpose of developing a 1sttime, 1st year student retention model to understand the determinant factors. Three years of data for 1st time, 1st year degree-seeking students at Western Kentucky University (USA). SAS Enterprise Miner was used to implement an ensemble model consisting of logistic regression, decision trees, and neural networks using the default average method. Each of the algorithms were also evaluated independently. After performance comparison, decision trees were preferred, based on reasoning similar to Delen [16]. Risk factors identified were high school GPA and programme loads.

Further research by Bogard et al. (2013) [18] continued to use SAS R Enterprise Miner to develop a model to score university students based on their probability of enrolment and retention early. Starting in the fall of 2008, the Western Kentucky University (USA) began warehousing applicant data files on a weekly basis. These data snaps have provided a rich data source for decision support and predictive modelling. With regard to retention, the concern was more with predictive accuracy than model clarity because, in previous work, student attrition factors were already addressed. In this research, they compared the results of four different algorithms, namely: neural networks, decision trees, gradient boosting and double scoring. Results presented problems as different evaluation measures ranked the methods differently.

The application of Data Mining and other Analytics into the educational context has increased in the last decade. Ferguson presents in [18] three drivers for this to occur: first, the volumes of data that are collected in educational institutions have greatly augmented, whether from Course or Learning Management Systems or Student Information Systems; second, the use of e-learning: although have helped collecting data it also have brought some learning issues such as possible lack of motivation and difficulties for the educators to receive direct feedback regarding the mood, level of interest, or even the understanding of the students; and finally, the political concerns: countries are getting more understanding about the importance of higher education for its development and have an interest to improve it, to offer better learning opportunities that lead to better academic results.

D. Kabakchieva compares in [16] Decision trees, a Bayesian classifier, a logistic model, a rule-based learner, and the Random Forest. They analyzed three different datasets are used to predict dropout in first-year Electrical Engineering students: Pre-university information, which is mainly the previous academic performance; the academic performance, i.e. the number of attempts of every course and the higher grade; and a combination of both. The results were very similar for those datasets including the grades data, which implies that the pre-university data does not add much independent information. Decision trees provide with good results between 75 and 90% of accuracy. It was necessary to implement Cost-sensitive learning in order to avoid False Negatives.

V. Conclusion

In this paper we review that existing techniques for Educational Data Mining to get acquire information as always levels of uncertainty in the planning of research and this proposal is no exception. It is planned though that it may possibly give the essential organization to structure the problem and put it in the context of existing research efforts and identify issues that remain open or where deep research contributions can be fulfilled.

References

[1] R.S.J.d. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions,"

Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–17, Oct. 2009.

[2] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 40, no. 6, pp. 601 –618, Nov. 2010.

[3] Krina Parmar, Prof. Dineshkumar Vaghela, Dr Priyanka Sharma, "Performance Prediction Of Students Using Distributed Data Mining" IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS'15, 2015.

[4] J. Ranjan and K. Malik, "Effective educational process: a data-mining approach," VINE: The Journal of Information and Knowledge Management Systems, vol. 37, no. 4, pp. 502– 515, 2007.

[5] J. P. Campbell and D. G. Oblinger, "Academic Analytics," Tech. Rep. October, 2007.

[6] R. A. Huebner, "A Survey of Education Data Mining Research," Research in Higher Education Journal, pp. 1–13, 2012.

[7] M. Bienkowski, M. Feng, and B. Means, "Enhancing teaching and learning through educational data mining and learning analytics: An issue brief," tech. rep., 2012.

[8] T. Hey and K. Tolle, The fourth paradigm data-intensive scientific discovery. Redmond, Wash.: Microsoft Research, 2009.

[9] S.B. Kelling, W. M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, and G. Hooker, "Data-intensive Science: A New Paradigm for Biodiversity Studies," BioScience, vol. 59, no. 7, pp. 613–620, Jul. 2009.

[10] S. Kotsiantis, "Supervised machine learning: A review of classification techniques," Informatica, vol. 31, pp. 249–268, 2007.

[11] V. Hodge and J. Austin, "A survey of outlier detection methodologies," Artificial Intelligence Review, vol. 22, pp. 85–126, 2004.

[12] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Systems with Applications, vol. 33, pp. 135–146, July 2007.

[13] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document is clustering techniques," KDD workshop on text mining, pp. 1–20, 2000.

[14] R.S.J.d. Baker, "Data mining for education," International Encyclopedia of Education, 2010.

[15] D. Delen, "A comparative analysis of machine learning techniques for student retention management," Decision Support Systems, vol. 49, pp. 498–506, Nov. 2010.

[16] D. Kabakchieva, "Student Performance Prediction by Using Data Mining Classification Algorithms," International Journal of Computer Science and Management Research, vol. 1, no. 4, pp. 686–690, 2012.

[17] M. Bogard and C. James, "Using SAS Enterprise BI and SAS Enterprise MinerTM to Reduce Student Attrition," tech. rep., 2012.

[18] M. Bogard, "A Data Driven Analytic Strategy for Increasing Yield and Retention at Western Kentucky University Using SAS Enterprise BI and SAS Enterprise Miner," tech. rep., 2013.