ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

Predictive Modeling in Educational Data Mining

Kapila Kundu[1], Ramkala Punia[2]

Research Scholar, Computer Science and Engineering Department, GJUS&T University, Hisar, India[1]

Research Scholar, Computer Science and Engineering Department, GJUS&T University, Hisar, India[2]

ABSTRACT: Educational Data Mining concerned with developing methods for exploring the educational data, and using those methods to better understand students' performance, and their learning environment. Predicting students' performance is dominant task in educational data mining that has always aroused the curiosity of researchers. This survey reviews the various studies carried out in predictive modeling in Educational Data Mining. This survey has been categorized on the basis methods used in predictive modeling. The main focus of this survey has been on the experimental data, methodology, conclusion, and future directions.

I. INTRODUCTION

Data mining involves extracting useful, interesting, interpretable and novel pattern from large amount of data [4]. Educational Data Mining (EDM) is a new field which combines data mining and education system. EDM concerned with developing methods for exploring the educational data, and using those methods to better understand students' performance, and their learning environment [5, 6, 7, 8, 9].

A variety of methods have been presented by the researchers in the field of educational data mining in recent years. Some methods are similar to basic data mining methods, which are used in other domains also like classification, clustering, association rule mining. On the other hand, some methods are unique to educational data mining like regression, ensemble methods [5, 6, 8, 9]. EDM involves many educational systems like traditional classroom, e-learning, learning management system, adaptive hypermedia educational systems, intelligent tutoring system, test/quizzes, and text/contents [7, 10, 11].

There are numerous applications or tasks that have been employed in educational environment. Romero (2010), Baker (2012) and Alejandro (2014) enlist many application areas and variety of functionalities related to students, teachers and about organization under the umbrella of EDM [5, 6, 12]. The most commonly used tasks by the researchers are as follows: student modeling, predicting students' behavior, assessment, feedback and support, teacher, curriculum and domain knowledge, and predicting students' performance [12]. These are the frequently used application areas in educational data mining. Predicting students' performance is dominant task in educational data mining that has always aroused the curiosity of researchers. Understanding factors that affect the students' performance is very important facet that helps in improving and understanding the educational landscape [5, 13]. By predicting student's performance, students can be helped at the early stages and this can improve the academic success of students. The main aim of this study is to investigate- i) the prediction methods in reference to forecasting the students' performance. ii) Clustering method to find the similar groups in students based on specific variables related to their demographic details and academic performance.

This research synopsis is organized as follows. The background details regarding predictive student's modeling are given in the next section. Literature review of the study is presented in section 3. The problem definition is given in section 4. The objectives of the study are included in section 5. The research methodology is presented in section 6. In section 7, future scope and relevance of the study is pointed out.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

II. STUDY AND REVIEW OF LITERATURE

Predicting students' performance means to detect the significant attributes, relationship between the attribute, and other aspects that affect students' performance. Many indicators help to detect the students' performance [40-43]. These are as follows: efficiency, evaluation, achievement, competence, resource consuming, elapsed time, correctness, deficiencies, etc. The main target of this task is to assess the learner if he/she will be able to accomplish a given task or reach a specific learning goal [12]. Researches carried out in predicting students' performance include different facets. These are as follows:

1. Predicting correlation between attributes
2. Predicting failure rate
3. Predicting drop-outs
4. Discovering significant attributes that influence the academic performance or the associated risks like failure or dropping out
5. Predicting unusual occurrences

This section is organized as follows: first section includes research papers in which classification is applied for predicting student's performance. Second section involves those papers in which regression is employed for predicting student's performance. Clustering related papers are included in last section.

II.i. CLASSIFICATION FOR PREDICTING STUDENTS' PERFORMANCE

Using classification, researchers derive significant result from the educational data. It is frequently used by researchers in educational data mining. Concerning Wang and Liao (2011), they have identified the performance of students to predict future performance in learning English as a second language. A back-propagation (BP) algorithm is used for the supervised cluster classification of student characteristics and learning performances. They have implemented three different levels of teaching content for vocabulary, grammar, and reading for adaptive learning in the AL-TESL-e-learning system. The feasibility of the AL-TESL-e-learning system is evaluated by comparing a control group in regular online course and an experimental group enrolled in the AL-TESL-e-learning system [40].

A very promising approach explored by Zafra, Romero, and Ventura (2011) based on multiple instances learning to predict students' performance and to improve the obtained results using single instance learning. The study uses the students' information from the virtual learning environment at Cordoba University using Moodle Platform. Algorithms using multiple instance learning representation show significantly better results with a confidence of 95% by Wilcoxon rank sum test [41].

Pardos, Wang, and trivedi (2012) have provided an analysis of the different impact on student test score prediction. They have employed linear regression and random forest. They have also included a K-mean clustering technique to improve prediction accuracy of algorithm on the same dataset. They have investigated the significance of performance prediction in the context of test score prediction and within-tutor knowledge inference. They have also raised the issue of leakage in prediction evaluation and its role in cross-validation accuracy [42].

Dorina (2013) have applied classification methods on the Bulgarian university sample data to reveal new patterns. They have found that prediction rates are not remarkable. They have also found the most influencing attribute in classification process. Harwati et al., (2015) focused on mapping students using K-mean Cluster algorithm to reveal the hidden pattern and classified students based on their demographic and average of course attending. They also considered the problem of class imbalance and encountered the effect of rebalancing of data on the algorithm [45].

Marbouti et al., (2016) have compared seven predictive methods to identify at risk students in a course. They have used feature selection method to reduce the number of variables. Out of seven methods, they created an ensemble method and the results showed that the ensemble method performed well and gave higher accuracy than the other methods [3]. Open University project analyzed and found at- risk students by using classifications method on the bases of demographic and virtual environment data [46]. Classification methods were also used to find students academic performance [47], to investigate courses that affect the performance of student [48], and also used to focus on the performance of instructor [49].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

II.ii. REGRESSION FOR PREDICTING STUDENTS' PERFORMANCE

Regression is another technique that has significant appearance in predicting students' performance. Many researchers have applied regression to predict continuous variable. Wang and Beck (2012), predict students' performance after a delay to determine if, and when, the student will retain the acquired knowledge. They use data for analysis from Assistments system for 4th through 10th grade mathematics. They built a logistic regression model, using the 27,468 data points with delayed practice opportunities. Main contributions of this work are as follows: first, the master learning notion is expanded to take into account the long-term effect of learning. The second contribution is extending the PFA (Performance Factors Analysis) model with features that likely to be relevant for retention. The last contribution is on discovering a new problem that is actionable by ITS (Intelligent Tutoring System) [42].

Olani (2009) has concluded that students' pre-college academic performance were strong predictors of GPA (Grade Point Average) at university level. The effects of psychological variables of university GPA are also not negligible. They have applied multiple regression analysis to identify the most important predictors of university GPA [50]. Marbouti et al., (2015) constructed three regression-based models to identify at-risk students in first year engineering course at three important time of the semester according to the academic calendar. The models were able to identify 79% of at-risk students at week 2, 90% at week 4, and 98% at week 9 [51].

A case study has been carried out by the Hussan and Al-Razgan (2016). They applied regression techniques on datasets of graduate and undergraduate students of Kingdom of Saudi Arabic (KSA) to find if the pre-university exams have a real effect on the students' college GPA. They found that high school GPA affects the college GPA more than pre-university exams, and that the enrolled year has an unexpected effect on the college GPA. They also investigated that the mean of students' college GPA is decreasing by time [52].

II.iii. CLUSTERING FOR PREDICTING STUDENTS' PERFORMANCE

Clustering is not so frequently used in predicting students' performance. Concerning Harwati, Ardita and Febriana (2016), they have focused on mapping students using K-mean clustering algorithm to reveal the hidden pattern and classifying students based on their demographic (gender, origin, GPA, grade of certain courses) and average of course attending [52].

Kerr and d'Chung (2012) have identified the key features of students' performance as a relevant step of the assessment cycle of evidence-centred design [53]. Mostow, Gonzalez-Brenes, and Tan (2011), have developed an Automatic Classifier of Relational Data system to perform the feature engineering by training classifiers directly on a relational database of events logged by a tutor. Their system has also learnt a classifier to predict whether a child finished reading a story in LISTEN's Reading Tutor [54].

III. CONCLUSION

Classification and regression are the two significant method used in predicting students' performance. Some researchers have applied classification and regression alone and some applied both the techniques in their studies. Clustering is another essential technique used in predicting students' performance. It is not used widely by the researchers but it is an important technique used to find the groups of students having similar properties.

Predicting students' performance will be helpful for the educational institutions in framing and implementing their policies for creating a quality learning environment. To be specific, the educational institutes can do resource planning, tune the study material to suit the needs of the students and take remedial actions to address the risks like dropping out from schools and failure rates. This study will contribute in creating learning and teaching environment that goes in tune with the need of the target groups in the students.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

REFERENCES

[1]. F. Marbouti, H.A. Diefes-Dux & K. Madhavan, " Models for early prediction of at-risk students in a course using standard-based grading", *Computer & Education*, vol. 103, pp. 1-15, 2016.
[2]. J. Han, M. Kamber, "*Data Mining: Concepts and Techniques"*, Morgan Kaufmann Publishers, 2006.
[3]. R. S. J. D. Baker, & K. Yacef, "The state of educational data mining in 2009: A review and future vision", *Journal of Educational Data Mining*, vol. 1, issue 1, pp. 1–15, 2009.
[4]. C. Romero, & S. Ventura, "Educational data mining: a review of the state of the art", *IEEE Transactions on Systems, Man, and Cybernetics, part C: Applications and Reviews*, vol. 40, issue 6, pp. 601–618, 2010.
[5]. C. Romero, & S. Ventura, "Educational data mining: a survey from 1995 to 2005", *Expert Systems with Applications*, vol. 33, issue 1, pp. 135–146, 2007.
[6]. C. Romero, S. Ventura, M. Pechenizkiy, & S. J. d. R. Baker (Eds.), "*Handbook of educational data mining, data mining and knowledge discovery series"*, Florida: Chapman & Hall/CRC, 2010.
[7]. R. S. J. D. Baker & P. S. Inventado, "Educational Data Mining and Learning Analytics", *Learning Analytics: from research to Practice*, pp. 61-75, 2014.
[8]. O. Zaïane, "Web usage mining for a better web-based learning environment", *In Proceedings of Conference on Advanced Technology for Education*, 2001, pp. 60–64.
[9]. C. Romero, & S. Ventura, "*Data mining in e-learning*", book series 4, WIT Press, 2006, pp. 328.
[10]. A. Pena-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works", *Expert Systems with Application*, vol. 41, pp. 1432-1462, 2014.
[11]. B Sen, E. Ucar & D. Delen, "Predicting and analyzing secondary education placement-test score: A data mining approach", *Expert Systems with Applications*, vol. 39, pp. 9468-9476, 2012.
[12]. S. D'Mello, & A. Graesser, "Mining bodily patterns of affective experience during learning" *In Proceedings of the 3rd international conference on educational data mining*, 2010, pp. 31–40.
[13]. R. Pelanek, "Metric for evaluation of student model", *Journal of Educational Data Mining*, vol. 7, issue. 2, 2015.
[14]. R. S. J. D. Baker, S. M. Gowda, M. Wixon, J. Kalka, A. Z. Wagner, A. Salvi, V. Aleven, G. W. Kusbit, J. Ocumpaugh, & L. Rossi, "Towards sensor-free affect detection in cognitive tutor algebra", *In Proceedings of the 5th international conference on educational data mining*, 2012, pp. 126–133.
[15]. Z. A. Pardos, Q. Y. Wang, & S. Trivedi, "The real world significance of performance prediction", *In Proceedings of the 5th international conference on educational data mining*, 2012, pp. 192–195.
[16]. Y. Wang, & J. E. Beck, "Using student modeling to estimate student knowledge retention" *In Proceedings of the 5th international conference on educational data mining*, 2012, pp. 200–203.
[17]. A.T. Corbett & J. R. Anderson, "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge", *User Modeling and User-Adapted Interaction*, vol. 4, pp. 253-278, 1995.
[18]. D. Kabakchieva, "Predicting student performance by using data minng methods for classification", *Cybernetics and Information Technologies*, vol. 13, issue 1, pp. 61-72, 2013.
[19]. J. Kuzilek, M. Hlosta, D. Herrmannova, Z. Zdrahal, & A. Wolff, "OU analyse: Analysing at-risk students at The Open University", *In Learning Analytics Review: LAK15*, 2015, pp. 1-16.
[20]. H. Hamsa, S. Indiradevi & J. J. Kizhakkethottam, " Student academic performance prediction model using decision tree and fuzzy genetic algorithm", *In Global colloquium in recent advancement and effectual Reserches in Engineering, Elsevier*, 2016, pp. 326-332.
[21]. Y. altujjar, W. Altamimi, I. Al-turaiki & M. Al-Razgan, "Predicting critical courses affecting students performance: A case study", *In Symposium on data mining application, Elsevier*, 2016, pp. 65-71.
[22]. A. M. Ahmed, A. Rizaner & A. H. Ulusoy, "Using data mining to predict instructor performance", *In 12th international conference on application of fuzzy systems and softcomputing, Elsevier*, 2016, pp. 137-142.
[23]. A. Olani, "Predicting first year university students' academic success", *Education & Psychology*, vol. 7, issue 3, pp. 1053-1072, 2009.
[24]. F. Marbouti, H. A.Diefes-Dux, & J. Strobel, "Building course-specific regression-based models to identify at-risk students", *In The american society for engineering educators annual conference*, 2015, pp. 2-11.
[25]. S. M. Hussan & M. S. Al-Razgan, "Pre-university exam effect on students GPA: A case study in IT", *In Symposium on data mining application, Elsevier*, 2016, pp. 127-131.
[26]. Harwati, A. P. Alfiani & F. A. Wulandari, "Mapping Student's Performance based on data mining approach", *In The international conference on agro-industry (ICoA): competitive and sustainable agro-industry for human welfare, Elsevier*, 2015, pp. 173-177.
[27]. D. Kerr, & G. K. W. K. d Chung, "Identifying key features of student performance in educational video games and simulations through cluster analysis", *Journal of Educational Data Mining*, vol. 4, issue 1, pp. 144–182, 2012.
[28]. J. Mostow, J. Gonzalez-Brenes, & B. H. Tan, "Learning classifiers from a relational database of tutor logs", *In Proceedings of the 4th international conference on educational data mining*, 2011, pp. 149–158.
[29]. A. M. Shahiri, W. Husain & N. A. Rashid, "A review on predicting student's performance using data mining techniques", *In The third information systems international conference, Elsevier*, 2015, pp. 414-422.