DOI: 10.15415/jotitt.2017.51002

# A Review on Prediction of Academic Performance of Students at-Risk Using Data Mining Techniques

**Preet Kamal[1], Sachin Ahuja[2]**

[1,2] CURIN, Chitkara University, Punjab, India.

E-mail: kamal.preet@chitkara.edu.in, sachin.ahuja@chitkara.edu.in

Published Online: June 28, 2017



**Abstract***:* Educational data mining is the procedure of converting raw data collected from educational databases into some useful information. It can be helpful in designing and answering research questions like performance prediction of students in academics, factors that affect the students' performance, help the teachers in understanding the problems faced by the students to understand the course content and complexity of the subject taken so that the teachers can take timely action to control the dropout rate. This also includes improving the teaching learning process so that the interventions can be taken at the right time to improve the performance of the student. This paper is the review of the research work done in the field of educational data mining for the prediction of students' performance. The factors that influence the performance of the students i.e. the type of classrooms they attend such as traditional or on-line, socio-economic, educational background of the family, attitude toward studies and challenges faced by the students during course progress. These factors leads to the categorization of the students into three groups "Low-Risk": who have High probability of succeeding, "Medium-Risk": who may succeed in their examination, "High-Risk": who have High probability of failing or drop-out. It elaborates the different ways to improve the teaching learning process by providing the students personal assistance, notes, class-assignments and special class tests. The most efficient techniques that are used in educational data mining are also reviewed such as; classification, regression, clustering and and prediction.

**Keywords:** Data Mining (DM), Educational data mining (EDM), Education System

Journal on Today's Ideas – Tomorrow's Technologies, Vol. 5, No. 1, June 2017 pp. 30–39

## INTRODUCTION

Educational Data Mining (EDM) is the application of Data Mining (DM) and its objective is to analyze the different types of data in order to resolve educational research issues [16]. Data Mining is the process of extracting useful and important information from data sets. It is being used by organizations, scientists and governments from last so many years to collect data like airline

passenger records and record of census data [2]. The volume of educational data has increased with advancement of technologies. It can be handled using Data Mining techniques. The educational institutes are also getting automated with the help of advanced technologies.

The educational research in Data Mining also contributes a lot to the predictive technologies. Data Mining is set up on the theory that the historic data retains the hidden and unknown information observed as a challenging task in data prediction. Data analysis is one way of forecasting the growth or decline in academic performance. The use of internet and e-learning in the field of education has facilitated the students. On the other hand offline education- is the medium to exchange knowledge and develop skills by face-to-face interaction. The tutor can easily understand the behavior of the student towards his studies. The data mining techniques can be applied to such data like students' behavior towards his studies, performance in their academics, family background and the data collected form students in classroom interactions. Such data help to create student models. E-learning and Learning Management System (LMS) is the combination of online instruction and communication that collaborates administration and reporting tools. Intelligent Tutoring (ITS) and Adaptive Educational Hypermedia System (AEHS) acquire background knowledge about teaching strategy and student behavior are few examples of student models. [17].

## EDUCATIONAL TASK

EDM plays an important role for every stakeholder. The following table explains the importance of EDM that helps in different ways for every stakeholder's.

Table I: Various Stakeholders

| | |
|---|---|
| Students | The activities that student involving in that affect their studies. Guide them at the right time about their short-falls and recommend them the relevant material that can help the students at various levels to improve their grades. |
| Mentors | To analyse the students on the basis of their performance and provide special attention to the students in case they score lower grades. Also the mentors can customize the teaching method according to the requirements of students. |
| Course planner | Considering the student performance the course planner should plan the course content according to the students' level of learning. Flexibility in the course content should be considered according to the changes in technology and learning ability of the students. |

The different data mining techniques that can be used to analyse the data available in the educational research issues related to student performance in their academics. Different systems have been created in various EDM research areas like E-Learning and Learning Management Systems(LMS), Adaptive Educational Hypermedia Systems(AEHS) and Intelligent Tutoring (ITS)[17]. These resources play an important role in the decision making of teachers about students' performance. So, that the required steps can be taken to improve the academic performance of the students. The performance of the students can also be assessed using previous data, such as: results drawn from class tests, assignments, seminars and assessments taken in class [4, 23]. The results collected from the previous data will help teachers as well as students to improve their academic record. Based on the marks scored by the students, they can be categorized in three groups: "Low-Risk": Who have High probability of succeeding. "Medium-Risk": Who may succeed in their examination, "High-Risk": Who have High probability of failing or drop-out. Intelligent Tutoring system used to Predict/Classify students into different categories and prediction about their results after applying the remedial actions [8, 9]. The performance of the students can be affected by various socio-economic factors such as family income of the students and their personal interest in the subject and, also the social distractions around students. [10, 52] and perceptions of the students about studies [11, 4]. Attendance also plays an important role in the performance of student as it predicts their attitude towards studies. The hours spent on self-study after the college is also an important factor to consider, educational background of the parents and their awareness about opportunities provided by the institutes nearby can play a vital role in the performance of student [19, 24, 25, 28].

The type of classroom attended by the student, affects the performance of the student i.e. traditional classroom or on-line education affects the learning ability of the student [19] [24]. Different strategies can be used to judge the students' reaction to a particular approach used to assess their performance. Data mining techniques used to detect the students attitude towards the assignments given and also the behavior towards their studies. Based on the different behaviours and personal habits of the students different groups of students can be formed such as: Hint-Driven (the students who can solve the problem with little assistance) or Failure-Driven (the students who cannot score marks even after taking help from the tutor or the person providing the assistance) and Low-Motivated students (the students who feel they cannot perform in the studies after every successful attempt). The required remedial actions can be taken that can lower down their drop-out rates [9] and also boost them to perform better. Two different data sources (logged interface and eye-tracking data) are used and discussed in exploratory learning environments in building

students models using data-based modelling framework in unsupervised and supervised learning environments [12].

To improve the performance of the students' various systems implemented that provide immediate feedback . ASSISTments have been considered to be one of the best systems. It uses the concept of scaffolding questions that break the problem into pieces, so that the student can solve the complex problems [10] hints are messages that provide insights and suggestions for solving a specific problem, and each hint sequence ends with a bottom-out hint which gives the required answer to the student. [1, 3, 8, 45, 51, 53]. To predict the student performance, these methods provide good results as the existing literature refers [28,30,35,36]. It is an online learning environment adopted by the open universities clubbing the students' data with their demographic data, by using VLE each student can be tracked individually and more focused support can be provided to the students. It highlights the students' pattern of behaviour that changes during the course as the complexity increases with the progress of the course. During the progression of the course five to seven assignments are provided to the students. Generally, the module ends by a final examination. The first assessment is considered to be a good predictor for the further performance of students. This predicts the right time of intervention with student so that the teacher can find out which students is at-risk and also offer them help at the right time. This can improve students' chances of success in academics. Online education platforms like Massive Open Online Courses (MOOCs), Coursera or Future-Learn are used to identify the students at-risk, they used predictive models using machine learning techniques. These models provide the information to the teachers about the performance of students. The four identified activity types that provide useful information for prediction, Resource: books and handouts for the students, Forum: the platform provided to the students to communicate with the tutors, Subpage: the means of navigating in the VLE environment, Content: specification and the guidelines. [39].

## TOOLS AND TECHNIQUES

To predict the students' performance and behaviour towards their studies, data mining provides various tools and technologies to extract the valuable information from the available data. The various models created using supervised and unsupervised learning are: Lean model, Item Response Theory Style, Rasch model [10, 48] GUHA and Markov Chain-based graphical models [35] Bayesian models [36].They are few examples of the models that help to predict the performance of students and also provide them the required guidance at the right time. Most popularly the classification techniques are

applied to the educational data to process relevant information [9, 12, 23, 25, 26, 30, 36, 39]. Other methods that can be applied on the data to create models using data mining techniques such as: Association rules, Clustering, Sequential pattern analysis. Association rules, Clustering, Classification, Sequential pattern analysis, decision trees, Bayesian network [1, 4, 10, 11, 12, 17]. Di . Discriminant analysis and neural networks are applied to obtain the efficiency of students [1, 2]. To conduct the study on factors that affects the performance of the students methods like classification trees and CHAID can be applied [19, 24].

The data mining tools came into existence as the need of the researchers to access the data also increased. The data mining tools play an important role in the filtering of data, as these tools provide various methods for processing of related data. With the advancement of technology, the volume of data is also increasing day by day in the field of education, business, algorithm development and applied research. The theoretical data needs to be converted into computerized data as the volume of data is also increasing. To process this digital data and to extract the useful information various commercial and open software are available in the market. WEKA tool is very popular amongst the researchers as compared to other tools as it is an open source tool and is JAVA based. The algorithm can be implemented using WEKA and the desired results can be obtained [6, 15, 20, 24, 49]process evaluation and optimization [6, 10]. RapidMiner is a powerful software platform that gives an integrated environment for machine learning, data mining, text mining and other business and prediction analysis. It gives good results with small data set. It classifies the testing data and used 10-fold validation method. [43, 46, 47].

The preferred techniques to apply on the data are: Neural network for Moodle logs, Pattern Mining for evaluating statistically relevant patterns for data where values need to be delivered, to build the trees for prediction C4.5 algorithms with ID3 proved better in terms efficiency. [20]. Sequential pattern mining is one of the unsupervised techniques of interest in acquiring the experience of student to use intelligent system and context free languages more expressive than regular [11]. Classification methods i.e. Rule Induction and Naïve Bayesian classifier used for the prediction of graduate students' grade. The students were clustered into groups using K-Means clustering algorithm [26, 42]. The concept of Scaffolding questions (breaks down problem into steps to eventually get the problem correct) to get the better results from students. Different models were implemented using these test results such as: Lean model, Item Response Theory Style, Rasch model [10]. The previous student database used to predict the division of student using classification techniques, from the number of approaches available for data classification, the decision

tree method preferred [41,44]. For the prediction of student performance at the end of semester the information like: assignment marks, attendance, and seminar and class test marks were collected [42]. This information can help both the teachers and students to raise their level of division. The students who need special attention were identified so that the ratio of failure in student can be controlled and the required action could be taken [23].

On the other hand, methods like decision-tree classification, support vector machine (SVM), general unary hypotheses automaton (GUHA), Bayesian networks, and linear and logistic regression have their own importance as compared to other methods. These methods used to build predictive models using data from several Open University (OU) modules. The Open University offers a good test-bed for this work. The authors have discussed how the predictive capacity of the different sources of data changes as the course progresses. It also highlights the importance of understanding how a students' pattern of behaviour changes during the course [29]. Methods like GUHA and Markov chain-based graphical models explored to provide useful insights into the students' behaviour during their studies. On the basis of these outcomes the teachers and supporting staff can plan the interventions and required steps to help them in proving their acadamic perfromance [35].

The latest work at the Open University which used data from VLE, clubbed with demographic data for the prediction of student failure or dropout. The methods and techniques like, machine learning techniques, Bayesian models, distance learning used to get the better results. The prediction combines the results of machine learning algorithms: 1. k Nearest Neighbours (k-NN): Demographic data. 2.Classification and Regression Tree (CART: VLE data. 3. Bayes network combines both demographic and VLE data used to find out the students at-risk their behaviour towards their study and performance in presenting their skills. CART and Bayes models are applied to the combined VLE and demographic data [36].

## LIMITATIONS AND FUTURE SCOPE

To improve the teaching learning process and to get the better results from students there is a need of personal interaction between student and teacher. It helps teacher to predict the students' performance and to better understand the problems faced by the students. The students can be assisted in a better way and their risk of failure can be minimised by providing them with the required interventions. This paper is the review of previous work done in the field of education using various data mining tools and techniques. By using these tools and techniques the factors derived by the researchers that affect the performance of the students are: class attendance, previous class grades

and socio-economic factors that influence the studies of the students. Based on these factors students were categorized: High-risk students, Medium-risk student and Low-risk students, Good students, Bad students, slow learners, fast learners etc. The classification techniques preferred by the researchers for prediction include: Neural Networks, Decision Tree, NaiveBayes models, k nearest-neighbour etc. Decision tree has much efficiency and accuracy as concluded by number of researchers. Most of the research work is done in the field of educational data mining is based on the online learning, the techniques and methods applied also focused on the online education. The traditional classroom has taken back seat in the research. There is a need to pay attention towards the traditional learning also. Methods that can be applied on traditional learning environment followed by majority of the Indian institutes are not explored properly.

## CONCLUSION

The experiments have been carried out in open universities in countries other than India. The research work done in this field is especially related to subjects like psychology, mathematics, history and home science. Very few focused on the technical courses like : computers. Majority of the studies conducted are predicting the performance of the students based on the demographic, academic and social factors alien to the Indian environment. Since Indian culture and living style is different, it demands a different study to exactly relate to its educational system. There is a need to further explore the Indian education system so that the factors that affect the students performance can be studied according to the Indian scenario.

## REFERENCES

1. Razzaq, Leena, and Neil T. Heffernan. “Scaffolding vs. hints in the Assistment System” International Conference on Intelligent Tutoring Systems. *Springer Berlin Heidelberg*, 2006.
2. Han, J., Kamber, M. (2006). Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher.
3. Feng, Mingyu, and Neil T. Heffernan. “Informing teachers live about student learning: Reporting in the assistment system” Technology Instruction Cognition and Learning 3.1/2 pp. 63 (2006).
4. Tahir, Syed, and S. R. Naqvi. “FACTORS AFFECTING STUDENTS’PERFORMANCE.” Bangladesh e-journal of sociology 3, no. 1 pp. 2 (2006).
5. Mierswa, Ingo, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. “Yale: Rapid prototyping for complex data mining tasks.” In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 935-940. ACM, 2006.
6. Nghe, Nguyen Thai, Paul Janecek, and Peter Haddawy. “A comparative analysis of techniques for predicting academic performance.” In 2007 37th Annual Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, pp. T2G-7. IEEE, 2007. http://dx.doi.org/10.1109/FIE.2007.4417993.

7. Superby, Juan-Francisco, J. P. Vandamme, and N. Meskens. "Determination of factors influencing the achievement of the first-year university students using data mining methods." In Workshop on Educational Data Mining, vol. 32, p. 234. 2006.
8. Romero, Cristóbal, Sebastián Ventura, Pedro G. Espejo, and César Hervás. "Data mining algorithms to classify students." In Educational Data Mining 2008.
9. Feng, Mingyu, Joseph Beck, Neil Heffernan, and Kenneth Koedinger. "Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standardized Test?." In Educational Data Mining 2008.
10. Antunes, Cláudia. "Acquiring background knowledge for intelligent tutoring systems." In Educational Data Mining 2008.
11. Amershi, Saleema, and Cristina Conati. "Combining Unsupervised and Supervised Classification to Build User Models for Exploratory." JEDM-Journal of Educational Data Mining 1, no. 1 pp.18-71 (2009).
12. Baker, Ryan SJD, and Kalina Yacef. "The state of educational data mining in 2009: A review and future visions." JEDM-Journal of Educational Data Mining 1, no. 1 pp. 3-17 (2009).
13. Ayers, Elizabeth, Rebecca Nugent, and Nema Dean. "A Comparison of Student Skill Knowledge Estimates." International Working Group on Educational Data Mining (2009).
14. Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11, no. 1 pp. 10-18 (2009).
15. Barnes, T., M. Desmarais, C. Romero, and S. Ventura. "Educational Data Mining 2009: 2nd International Conference on Educational Data Mining." Proceedings Cordoba, Spain (2009).
16. Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40, no. 6 pp. 601-618 (2010).
17. Razzaq, Leena, Jozsef Patvarczki, Shane F. Almeida, Manasi Vartak, Mingyu Feng, Neil T. Heffernan, and Kenneth R. Koedinger. "The Assistment Builder: Supporting the life cycle of tutoring system content creation." IEEE Transactions on Learning Technologies 2, no. 2 pp. 157-166. (2009). http://dx.doi.org/10.1109/TLT.2009.23.
18. Ramaswami, M., and R. Bhaskaran. "A CHAID based performance prediction model in educational data mining." arXiv preprint arXiv:1002.1144(2010).
19. Kumar, S. Anupama, and M. N. Vijayalakshmi. "Efficiency of decision trees in predicting students' academic performance." In First International Conference, on Computer Science, Engineering and Applications, CS and IT, vol. 2, pp. 335-343. (2011).
20. Kumar, Varun, and Anupama Chadha. "An empirical study of the applications of data mining techniques in higher education." *International Journal of Advanced Computer Science and Applications* 2, no. 3 (2011). http://dx.doi.org/10.14569/IJACSA.2011.020314.
21. Shih, Benjamin, Kenneth R. Koedinger, and Richard Scheines. "A response time model for bottom-out hints as worked examples." Handbook of educational data mining pp. 201-212 (2011).
22. Yadav, Surjeet Kumar, Brijesh Bharadwaj, and Saurabh Pal. "Data mining applications: A comparative study for predicting students' performance." arXiv preprint arXiv:1202.4815 (2012).
23. Yadav, Surjeet Kumar, Brijesh Bharadwaj, and Saurabh Pal. "Data mining applications: A comparative study for predicting students' performance." arXiv preprint arXiv:1202.4815 (2012).
24. Kabakchieva, Dorina. "Student performance prediction by using data mining classification algorithms." *International Journal of Computer Science and Management Research* 1, no. 4 pp. 686-690 (2012).
25. Tair, Mohammad M. Abu, and Alaa M. El-Halees. "Mining educational data to improve students' performance: a case study." *International Journal of Informational* 2, no.2 (2012).

26. Asiri, Mahdi, and Behrouz Minaei. "Predicting GPA and academic dismissal in LMS using educational data mining: A case mining." In 6th National and 3rd International conference of e-Learning and e-Teaching, pp. 53-58. IEEE, 2012.
27. Wolff, Annika, Zdenek Zdrahal, Andriy Nikolov, and Michal Pantucek. "Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment." In Proceedings of the third international conference on learning analytics and knowledge, pp. 145-149. ACM, 2013.
28. Demšar, Janez, and Blaž Zupan. "Orange: Data Mining Fruitful and Fun-A Historical Perspective." Informatica 37, no. 1 (2013).
29. Wolff, Annika, Zdenek Zdrahal, Drahomira Herrmannova, and Petr Knoth. "Predicting student performance from combined data sources." In Educational Data Mining, pp. 175-202. Springer International Publishing, 2014.
30. Pandey, Mrinal, and Vivek Kumar Sharma. "A decision tree algorithm pertaining to the student performance analysis and prediction." *International Journal of Computer Applications* 61, no. 13 (2013).
31. Huang, Shaobo, and Ning Fang. "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models." Computers & Education 61 pp. 133-145. (2013). https://doi.org/10.1016/j.compedu.2012.08.015.
32. Romero, Cristóbal, Manuel-Ignacio López, Jose-María Luna, and Sebastián Ventura. "Predicting students' final performance from participation in on-line discussion forums." Computers & Education 68 pp. 458-472. (2013). https://doi.org/10.1016/j.compedu.2013.06.009.
33. Marquez-Vera, Carlos, Cristóbal Romero Morales, and Sebastián Ventura Soto. "Predicting school failure and dropout by using data mining techniques." IEEE Revista Iberoamericana de Tecnologias del Aprendizaje8, no. 1 pp. 7-14. (2013). https://doi.org/10.1109/RITA.2013.2244695.
34. Hlosta, Martin, Drahomira Herrmannova, Lucie Vachova, Jakub Kuzilek, Zdenek Zdrahal, and Annika Wolff. "Modelling student online behaviour in a virtual learning environment." (2014).
35. Wolff, Annika, Zdenek Zdrahal, Drahomira Herrmannova, Jakub Kuzilek, and Martin Hlosta. "Developing predictive models for early detection of at-risk students on distance learning modules." (2014).
36. Patil, Priyanka Anandrao, and R. V. Mane. "Prediction of Students Performance Using Frequent Pattern Tree." In Computational Intelligence and Communication Networks (CICN), 2014 International Conference on, pp. 1078-1082. IEEE, 2014.
37. Ahmed, Abeer Badr El Din, and Ibrahim Sayed Elaraby. "Data Mining: A prediction for Students' Performance Using Classification Method." *World Journal of Computer Application and Technology* 2, no. 2 pp. 43-47 (2014).
38. Kuzilek, Jakub, Martin Hlosta, Drahomira Herrmannova, Zdenek Zdrahal, and Annika Wolff. "OU Analyse: analysing at-risk students at The Open University." Learning Analytics Review pp. 1-16 (2015).
39. Ahmad, Fadhilah, Nur Hafieza Ismail, and Azwa Abdul Aziz. "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques." Applied Mathematical Sciences 9, no.129 pp. 6415-6426. (2015).
40. Saxena, Ritika. "Educational data Mining: Performance Evaluation of Decision Tree and Clustering Techniques using WEKA Platform. "" *International Journal of Computer Science and Business Informatics* (2015).
41. Deshpande, Akshay, Prashant Pimpare, Shashank Bhujbal, Abhishek Kommwar, and Jagruti Wagh. "Student Performance Analysis, Visualization and Prediction Using Data Mining Techniques." *Imperial Journal of Interdisciplinary Research* 2, no. 5 (2016).
42. Puyalnithi, Thendral, V. Madhu Viswanatham, and Ashmeet Singh. "Comparison of Performance of Various Data Classification Algorithms with Ensemble Methods Using RAPID-

MINER." *International Journal* 6, no. 5 (2016).
43. Rana, Shiwani, and Roopali Garg. "Evaluation of Students' Performance of an Institute Using Clustering Algorithms." International Journal of Applied Engineering Research 11, no. 5 pp. 3605-3609. (2016).
44. Ostrow, Korinn S., and Neil T. Heffernan. "Studying Learning at Scale with the ASSISTments TestBed." In Proceedings of the Third (2016) ACM Conference on Learning@ Scale, pp. 333-334. ACM, 2016.
45. Asif, R., Haider, N. G., & Ali, S. A. (2016). Prediction of UndergraduateStudents' Performance using Data Mining Methods. *International Journal of Computer Science and Information Security*, 14(5), 374.
46. Kumar, M., Shambhu, S., & Aggarwal, P. (2016). Recognition of Slow Learners Using Classification Data Mining Techniques. *Imperial Journal of Interdisciplinary Research*, 2(12).
47. Ferrara, S., & Way, D. (2016). 2 Design and Development of End-of-Course Tests for Student Assessment and Teacher Evaluation. In Meeting the Challenges to Measurement in an Era of Accountability (p. 11). Routledge.
48. Alcala-Fdez, J., Garcia, S., Fernandez, A., Luengo, J., Gonzalez, S., Saez, J. A., ... & Herrera, F. (2016). Comparison of KEEL versus open source Data Mining tools: Knime and Weka software.
49. Read, J., Reutemann, P., Pfahringer, B., & Holmes, G. (2016). Meka: a multi-label/multi-target extension to weka. *Journal of Machine Learning Research*, 17(21), 1-5.
50. Ostrow, K. S., & Heffernan, N. T. (2016, April). Studying Learning at Scale with the ASSISTments TestBed. In Proceedings of the Third (2016) ACM Conference on Learning@ Scale (pp. 333-334). ACM.
51. Atinaf, W., & Petros, P. (2016). Socio Economic Factors Affecting Female Students Academic Performance at Higher Education. Health Care: Current Reviews, 4(163), 2.
52. Feng, M., & Roschelle, J. (2016, April). Predicting Students' Standardized Test Scores Using Online Homework. In Proceedings of the Third (2016) ACM Conference on Learning@ Scale (pp. 213-216). ACM.