IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 2, February 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002054

Literature Survey on Predicting Students Performance Using Data Mining and Machine Learning Techniques

A.Gokila Devi[1*], Dr.T.Christopher[2]

PG and Research, Department of Computer Science, Government Arts College Coimbatore (Autonomous), Tamil Nadu, India[1]

Assistant Professor, PG and Research Department of Information Technology, Government Arts College Coimbatore (Autonomous), Tamil Nadu, India[2]

ABSTRACT: Presently,predicting student's performance is a significantpart in higher educational institutions. The normsfor greatvalue universities aredepending on its outstandingeducational successes.The advancement in the data mining field makes it possible to mine educational data for enhancing the excellence of the educational processes. Accurate prediction of student activities at prior stage supports in recognizing weak students and enablesschools or colleges to provide helpfulsolutions to avoid them failure. Currently, several techniques are designed for forecasting student's activities. Recently, Educational Data Mining (EDM)is broadly used in educational area for extracting helpful details and hidden attributes in large-scale academic corpus. It is applied to estimate the student's activities. So, it might support the professorsto offer an efficient learning schemes. This article presents a detailed survey of various schemes proposed for student performance prediction. Also, a comparative analysis is presented to know the limitations of each scheme and provide a suggestion for further improvement in student performance prediction effectively.

KEYWORDS: Data mining,EDM, student activity forecasting, efficient teaching.

I. INTRODUCTION

Schooling is a crucial problem about the growth of a nation. The main intention of higher education institution is to present quality education to its students. Students are the only source of information about the learning environment and they are the only ones who can rate the effectiveness, the quality, the satisfaction of method of instruction, homework, textbook and course content. A method to accomplish the higher range of excellence in high-level academic is via forecasting student's educational activities and so enchantingprior solutions for enhancing student's activities and learning superiority.Student's activity estimation [1]is also act as a part of staff appraisal system. The relevant knowledge is hidden by the academic details and extractable via data mining approaches. If the data mining approaches are applied in the academic field, then it is called as EDM[2].

In EDM, estimating student activities [3-5] is taken as an important task. This process may offer students primaryresponsesfor supporting them to enhance their learning outcomes. A worthyconsistent framework which precisely forecasts student's activities can alternate the recent generalized studies, therefore minimizing the stress on schooling and knowledgefor studiesand also valid duration and strength for trainers and students. Specifically, the student's activity estimation is the procedure that how students train, and rapidly or graduallythey familiarize to current issues or if it is probable to gather the data demands for solving these issuesunswervingly from student's activities details [6].

Investigations in this EDM is conducted at a coachingstage, course stage or degree stage and so on [7-8]. The purpose of EDM [9] is mainly focused on the growth of advanced schemes for extracting the primary attributes from students' documents and employ them to know the students' skills and forecast their upcoming activities and successes.This

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002054

paper overviews the previous researchers related to the prediction of student's activities via different approaches. The main intension of a paper is studying in detailed information on different approaches for student performance prediction. Also, the limitations of those approaches are addressed to further improve the accuracy of estimating the student's activities effectively.

The following sections in this paper is prepared as: Section II provides the previous researches related to student performance prediction using various data mining approaches. Section III compares the performance efficiency of those approaches and algorithms and Section IV concludes the survey that reviews an entire discussion.

II. PREDICTING STUDENT PERFORMANCE

Multi-Relational Factorization model [10]was proposed for estimating the student's activities. Especially this model was exploited multiple relationships between tasks, students, and their Meta data by using Multi-Relational Matrix Factorization (MRMF). However, this method treated each correlation equivalently. So,a Weighted MRMF (WMRMF) was proposed for considering the significance of main relation. A model [11] was presented for estimating weak activities of student'sat a particular registration by Decision Tree (DT) and Naïve Bayes (NB) classifier. The presented model intention was to predict attrition in the student's registrations. Previous educational files and details from the registration were applied for training DT and NB.

Novel techniques [12] were developed for estimating the efficiency of Mass Open Online Courses (MOOCs). It studied the student performance with an objective of predicting whether auser could be Correct on First Attempt (CFA) in responding a query. The developed techniques controlactivity information gathered via MOOCs. A video-watching clickstream data was applied to extract summary quantities for every user-video set. So, the developed technique weredesigned using Kernel Density Estimator (KDE)for computing perfect duration from training data and using the respective victorymeasures as training attributes.

To analyzehistorical student course grade data, the utilization of knowledge technique [13] was proposed which predicted the student's grade. Logistic regressions were utilized in this proposed technique for classification process. It was depending on the marks obtained by every student in each classical program. It supported actual assumption that was depending on the concept that marks were utilized for estimating the student's dropout. Moreover historical drop out models were applied for detecting potential dropouts.

A Grade prediction algorithm [14] has been suggested to personalize and timely predict every student's mark. It addressed achallenge of predicting every student's resultant mark individually. It was learned for estimating the student's overall grade according to the information from classes held in previous yearsand results. At a certain time, residual and the total grade were predicted.The conventional machine learning algorithms [15] were utilized for creating a predictive framework of Grade Point Average (GPA) according to the group of self-regulatory training process. It quantified the detectability of every constituentsof the created framework and its relevance. A cross validation procedure external and Maximum Weighted Dependence Trees (MWDT) were considered to select and classify the features, respectively.

An enhanced method [16]has been developed for forecasting student performance. The proposed approach was used tensor factorization to forecast the performance of student.The estimation of every student was personalized and used to suggest the actions to the students. In this method, three mode tensor factorization was used. By applying moving average approach with a particular time, student's activity in a certain program was estimated.

For designing a student's educational activities recognition, DT and fuzzy Genetic Algorithm (GA)[17]were introduced. At first, the features like sessional scores, internal scores and registration value of Bachelor and Master degree students in Computer Science and Electronics and Communication courses was collected. In these features, internal mark is the mixture of assignment scores, mean scores and attendance scores acquired from 2 sessional

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002054

examinations. The SSLC and HSCscores and entrance scores, the admission score was obtained. The collected features were given as input to DT and fuzzy GA which predict the student's activities.

For prediction of student's academic performance, an integrated multiple classifiers[18] were proposed. It has an Aggregating One- Dependence Estimators (AODE), DT and K- Nearest Neighbor. These classifiers were integrated by using a likelihood for estimating the educational activities of engineering students. Each classifier generated hypotheses and every output classes of each classifier a posterior probability was generated and it were multipliedfor obtaining the likelihoods and the class was represented by maximum of posterior probability. It was chosen to be voting hypothesis for the final decision.

A new deep learning based algorithm called GritNet [19] was proposed for estimating the student's activities. It was built upon Bi-directional Long Short Term Memory (BLSTM) and GMP layers. The student's raw event records were given as input to the GridNetfor encoding the time stamped logs into a sequence of fixed length input vectors.Ensemble-based Progressive Prediction (EPP)[20] was developed for estimating the student's activities in degree courses. Based on states of students evolving performance, a bi-layered design involving many base estimators and a cascade of ensemble estimators was developed. Then a data-driven mechanism according to the latent factor and probabilistic MF was applied for finding course similarity to construct effective base estimators.

A fuzzy C-means clustering algorithm [21] was presented using 2D and 3D clustering for analyzing the student's performance depending on their examination results from college of computer science and technology, Huaqiao University for students registered in 2014. An optimized mining algorithm [22] was developed to analyze student's learning degree based on the dynamic data. Initially, the optimized text classification was used for matching the query texts to the knowledge points in an automatic way. After, the subjective weighting scheme was integrated with the expert knowledge for creating the training level matrix of students on knowledge records. At last, the DBSCAN clustering was used to group the personalized learning characteristics of students depending on the learning degree matrix.

An improved method [23] was designed using student's assignment submission behavior for predicting the activity of students with learning complexities via procrastination behavior. First, the feature vectors were constructed for defining the submission behavior of students for every assignment. After, a clustering was performed for labeling students as a procrastinator, procrastination candidate or non-procrastinator.

An attribute extraction scheme [24] was introduced depending on the student's knowledge details. After, the optimized Support Vector Regression (SVR) was applied to estimate the student's dropout. Here, 3 variables of SVR were not randomly chosen but computed via an enhanced quantum Particle Swarm Optimization (PSO).

III. RESULTS AND DISCUSSIONS

This section presents a detail about merits and demerits of different approaches used for predicting the student's academic activities studied in the previous section. The following challenges are addressed in these approaches. A high amount of information correlation is required in WMRMFthat highly reflects the task's difficulty. The accuracy of NBis degraded after considering the educational details of the second registration. The RMSE and AUC values of kernel density estimator are roughly identical to the conventional methods.

The logistic regression model is not applicable for non-numeric predicting variables. The efficacy of rank estimation steeply increases only after the mid-term and final exams. The MWDTs needs improvement on precision. The tensor factorization and moving average approach has high training and testing time. The DT and fuzzy genetic method has less efficiency problem. In GridNet, accurate estimation of student's activities is still challenging for first few weeks. Each term becomes much more complicated in ensemble learning. Through Table 1, the limitations in predicting student's performance are addressed.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002054

Table 1. Comparative Analysis of Student Performance Prediction Schemes

Ref.No.	Schemes Used	Benefits	Drawbacks	Dataset Used	Evaluation Metrics
[10]	MRMF, WMRMF	Better RMSE	Needs a high amount of information correlations which highly reflect the task's difficulty	KDD [25] Challenge 2010, ASSISTments Platform	**Root Mean Squared Error (RMSE)** Algebra: MRMF = 0.308 WMRMF = 0.3 Bridge: MRMF = 0.297 WMRMF = 0.296 Assistments: MRMF = 0.47 WMRMF = 0.458
[11]	DT, NB	Can estimate the loss of educationaldata	NB performance is degraded after including educational information of the second registration	10-fold cross validation	**Balance Accuracy** NB= 85% DT= 80%
[12]	Kernel Density Estimator	Improves prediction quality	RMSE and AUC values are roughly identical to the conventional methods	libFM	Accuracy Improvement= 9.52% RMSE Improvement = 14.48% Area Under Curve (AUC) Improvement = 51.85%
[13]	Logistic regression model	Reduce student dropout ratio	Not applicable for non-numeric predicting variables	Historical record of 5 varied BS in Computer Engineering programs	Precision = 98.95% Recall = 96.73% Specificity = 97.14% Accuracy = 97.13%
[14]	Grade Prediction Algorithm	Less prediction error	Less effectiveness	undergraduate digital signal processing course over the previous 7 years	Mean Absolute Error = 0.7
[15]	Cross validation procedure	Proposed intervention	Needs improvement on efficiency	Customized eighteen questions	Accuracy = 82%

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002054

	external, Maximum Weighted Dependence Trees	was simply handled by students			
[16]	Tensor factorization, Moving average approach	Minimize the memory consumption and human effort	High training and testing time	Algebra-2008-2009 train and test, Bridge-to-Algebra-2008-2009 train and test	**Algebra** RMSE = 0.30159 Training time = 908.71 mins Testing Time = 15.11 secs **Bridge** RMSE = 0.28700 Training time = 466.01 mins Testing Time = 6.61 secs
[17]	DT, Fuzzy GA	Provide prior action for enhancing student performance	Less efficiency	Student record from Bachelor and Master degree in Computer Science and Electronics and Communication courses	**Mathematics** **DT** Prediction Result (safe) = 50 Prediction Result (At Risk) = 50 **Fuzzy GA** Prediction Result (safe) = 80 Prediction Result (At Risk) = 20 **Mathematics** **DT** Prediction Result (safe) = 30 Prediction Result (At Risk) = 70 **Fuzzy GA** **Big data Result** Prediction Result (safe) = 80 Prediction Result (At Risk) = 20
[18]	DT, K- Nearest Neighbor and Aggregating One-Dependence Estimators	High efficiency for filtered dataset	DT has less accuracy for one dataset	Dataset is formed from an engineering college in India	Prediction accuracy (filtered dataset) = 98.86
[19]	GritNet	Improve student graduation prediction	For first few weeks accurate predictions are most challenging	ND-A and ND-B dataset	Mean AUC (ND-A dataset) = 92% Mean AUC (ND-B dataset) = 96%

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002054

		accuracy			
[20]	Bi-layered design using many base estimators, Cascade of ensemble estimators	Best prediction performance	Ensemble learning in each term becomes much more complicated	Student record from the Mechanical and Aerospace Engineering department at UCLA	MSE = 0.22 (@12 time)
[21]	Fuzzy C-means clustering	Good performance	It considers only student's grades whereas other types of student's details were needed to improve the accuracy	Examination results from college of computer science and technology, Huaqiao University for students registered in 2014	-Nil-
[22]	Optimized mining algorithm, DBSCAN	Can deal with large-scale dataset effectively	It depends on expert knowledge and considers only dynamic data	Dataset collected for testing online in a province of China in January 2019	-Nil-
[23]	Clustering algorithm	Simple and good accuracy	It considers the fixed length feature vectors	Student data from the University of Tartu in Estonia	Accuracy=86%
[24]	Feature extraction, SVR and improved quantum PSO	Betteraccuracy	The overall average running time was high	KDD Cup 2015 dataset	Accuracy=0.94 Overall average running time=2.1415sec

IV. CONCLUSION

This paper presents a detailed review of predicting the student'sactivities based on various data mining approaches. It addresses that various schemes have been encountered for recognizing student's academic activities effectively. These discussed approaches provide the recent developments in the prediction of student's activitiesvia enhanced solutions. According to thisinvestigation, EPP schemeattains improved efficiency compared to all other schemes. This survey helps to derive the motivation for our upcoming researches as well.

REFERENCES

[1] Agarwal, H., &Mavani, H. (2015). Student performance prediction using machine learning.*International Journal of Engineering Research and Technology*, *4*(3), 111-11.

[2] Jacob, J., Jha, K., Kotak, P., &Puthran, S. (2015). Educational data mining techniques and their applications. In *IEEE International Conference on Green Computing and Internet of Things*,pp. 1344-1348.

[3] Ramesh, V. A. M. A. N. A. N., Parkavi, P., &Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. *International Journal of Computer Applications*, *63*(8), 35-39.

[4] Ip, H. H. S., Li, C., Leoni, S., Chen, Y., Ma, K. F., Wong, C. H. T., & Li, Q. (2018). Design and evaluate immersive learning experience for massive open online courses (MOOCs). *IEEE Transactions on Learning Technologies*, *12*(4), 503-515.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002054

[5] Cantabella, M., Martínez-España, R., Ayuso, B., Yáñez, J. A., & Muñoz, A. (2019). Analysis of student behavior in learning management systems through a big data framework. *Future Generation Computer Systems*, *90*, 262-272.

[6] Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: a review. *Applied Sciences*, *10*(3), 1042.

[7] Ang, K. L. M., Ge, F. L., & Seng, K. P. (2020). Big educational data & analytics: survey, architecture and challenges. *IEEE Access*, *8*, 116392-116414.

[8] Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, *8*, 55462-55470.

[9] Dutt, A., Ismail, M. A., &Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, *5*, 15991-16005.

[10] Thai-Nghe, N., Drumond, L., Horváth, T., & Schmidt-Thieme, L. (2011). Multi-relational factorization models for predicting student performance. In *Proceedings of the KDD Workshop on Knowledge Discovery in Educational Data*, pp. 27-40.

[11] Guarín, C. E. L., Guzmán, E. L., & González, F. A. (2015). A model to predict low academic performance at a specific enrollment using data mining. *IEEE RevistaIberoamericana de Tecnologias del Aprendizaje*, *10*(3), 119-125.

[12] Brinton, C. G., & Chiang, M. (2015). MOOC performance prediction via clickstream data and social learning networks. In *IEEE Conference on Computer Communications*,pp. 2299-2307.

[13] Burgos, C., Campanario, M. L., de la Pena, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: a tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, *66*, 541-556.

[14] Meier, Y., Xu, J., Atan, O., & Van Der Schaar, M. (2015, November). Personalized grade prediction: A data mining approach. In *IEEE International Conference on Data Mining*,pp. 907-912.

[15] Zollanvari, A., Kizilirmak, R. C., Kho, Y. H., & Hernández-Torrano, D. (2017). Predicting students' GPA and developing intervention strategies based on self-regulatory learning behaviors. *IEEE Access*, *5*, 23792-23802.

[16] Thai-Nghe, N., Horváth, T., & Schmidt-Thieme, L. (2011). Factorization Models for Forecasting Student Performance. In *EDM* (pp. 11-20).

[17] Hamsa, H., Indiradevi, S., &Kizhakkethottam, J. J. (2016). Student academic performance prediction model using decision tree and fuzzy genetic algorithm. *Procedia Technology*, *25*, 326-332.

[18] Pandey, M., &Taruna, S. (2016). Towards the integration of multiple classifier pertaining to the Student's performance prediction. *Perspectives in Science*, *8*, 364-366.

[19] Kim, B. H., Vizitei, E., & Ganapathi, V. (2018). GritNet: student performance prediction with deep learning. *arXiv preprint arXiv:1804.07405*.

[20] Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, *11*(5), 742-753.

[21] Li, Y., Gou, J., & Fan, Z. (2019). Educational data mining for students' performance based on fuzzy C-means clustering. *The Journal of Engineering*, *2019*(11), 8245-8250.

[22] Shao, Z., Sun, H., Wang, X., & Sun, Z. (2020). An optimized mining algorithm for analyzing students' learning degree based on dynamic data. *IEEE Access*, *8*, 113543-113556.

[23] Hooshyar, D., Pedaste, M., & Yang, Y. (2020). Mining educational data to predict students' performance through procrastination behavior. *Entropy*, *22*(1), 12.

[24] Jin, C. (2020). MOOC student dropout prediction model based on learning behavior features and parameter optimization. *Interactive Learning Environments*, 1-19.

[25] KDD Cup, "Educational data minding challenge," https://pslcdatashop: web:cmu:edu/KDDCup/, 2010.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details