Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395767

A review on machine learning based student's academic performance prediction systems

Rahul Katarya
Computer Science Department
Delhi Technological University
New Delhi, India
rahuldtu@gmail.com

Jalaj Gaba
Computer Science Department
Delhi Technological University
New Delhi, India
jalajgaba12345@gmail.com

Aryan Garg
Computer Science Department
Delhi Technological University
New Delhi, India
aryangarg811@gmail.com

Varsha Verma
Computer Science Department
Delhi Technological University
New Delhi, India
varshaverma53@gmail.com

***Abstract*—Prediction of academic performance of students beforehand provides scope to universities to lower their dropout rate and help the students in improving their performance. In this field, research is being done to find out which algorithm is best to use and which features should be considered while predicting the academic performance of students. This kind of research work has been increasing over the years. This paper performs a survey on the techniques used in various research papers for academic performance prediction and also point out the limitations if any, in the methodology used.**

Keywords—Classification; Data Mining; Machine Learning; Prediction; Performance

I. INTRODUCTION

Education is the process of facilitating learning and acquiring skills, concepts, knowledge for the mental, physical, and spiritual development of an individual. An educated person is respected everywhere in the world and leads to the betterment of the society and prosperity of his nation. More importantly, it is necessary for his personal development.

Academics play a crucial part in education, and a student's academic performance in colleges and high school is an important criterion to judge a student's success. The academic performance comprises, in most cases, the grade point in each course studied and the cumulative grade point for the entire year (also known as CGPA). In the modern education system, it is the best and most suitable way known to judge one's caliber in academics. The grades that a student achieves greatly influence his chances of getting good colleges for higher studies and his job.

In the past few decades, there have been many attempts to get a prediction of a student's academic performance before the starting of his course, so that results can be predicted. This also becomes necessary as a student may like to know the areas where he lacks so that he can improve. Timely prediction of the future result may help an instructor know on which student to focus in particular and in what areas the improvement might be needed. An institute or the government would also like to know it to check the efficiency of its current education system and to make it better in the longer run.

Sachio Hirokava showed that academic performance depends on various factors such as demographic, behavioral, past results, habits, etc. [1] Sometimes unexpected factors like the place a student is living in might also produce a significant impact on it. There have been many kinds of research going on to tap these effects and the importance of different variables on a student's academic performance. These factors may not be the same globally, and each university and school may have individual differences when it comes to them.

In today's world, data has become very powerful, and Machine Learning has become very useful in harnessing the power of that data. Deep learning (DL) can be seen as an essential component of Machine learning which uses a computer model based on neural networks.[2]

Machine learning, along with deep learning techniques, has played a very pivotal part in the prediction of student's academic performance. There have been various Machine learning and deep learning methods like SVM, KNN, clustering, etc. that have been studied on various datasets belonging to different institutions to find hidden and unexpected patterns. Some new machine learning methods have also been developed and applied.

This paper has surveyed and analyzed the various work done in this field, and a comprehensive study is presented. The aim is to study the best variables used in prediction and the best

978-1-7281-9537-7/21/$31.00 ©2021 IEEE

algorithms so further work can be done to improve the existing models and create a general system for prediction which can be globally applicable and is accurate.

This paper is divided as- Section II consists of the literature that we have reviewed along with some specific findings. Section III contains our contributions and Section IV contains the conclusion with the suggestion of future work followed by the references.

II. Literature Review

Students' CGPA is predicted by using ANN. The data used include the social and economic background of students. It also includes competitive exam results. Factors that impact students' performance are found out by statistical evaluations. The neural network has 11 input variables. Total Neuron layers- 2 hidden and 1 output [3]. One limitation is that attributes are not taken holistically. Other attributes like extra-curricular activities, online classes/courses can also be considered.

It was predicting at-risk students at an early stage using SGPA and CGPA to decrease the chance of student failure. It predicts the minimum required GPA for a particular student for each semester [4]. It helps in early identification, constant monitoring, and timely remedial action.

Students' scores are predicted using the Gartner Analytics Ascendancy model. The data used includes marks scored by students previously in a particular STEM course. It makes use of 8 Machine Learning algorithms in total. A student's academic performance is predicted in a particular course. Results are used to improve STEM education by developing appropriate tools [5]. The Linear regression algorithm predicted scores with the best accuracy, whereas the Naive Bayes classifier predicted with worst accuracy.

In [6], a novel framework MTLTR-APP is developed to predict students' academic performance. The multi-task learning framework is used. Behavioral data of students is used for prediction like sleep patterns.

Support Vector Machines are used to segregate those students who have low performance in academics. Students with the risk of failures are accurately identified. It uses SVM with a radial basis function kernel for training our model for each course [7]. It uses a dataset from 2 universities- University A (10 weeks courses), University B(12 weeks course).

Various Machine learning techniques are used to find out whether course grades of individual subjects or GPA should be used for predicting academic performance. By looking at the results it can be said that course grades should be used until the third term. After that, the GPA should be used [8]. This helps to determine which is to be used when.

Students' academic performance is predicted using a Multi-label ensemble model. It makes use of data-mining to improve students' academic performance. It predicts the performance of students in the next semester for different courses using different machine learning algorithms like SVM, RF, KNN, MLP to train the model. Label Powerset is used to further improve the prediction accuracy [9].

Predicting student's performance before the commencement of a course [10]. SVM, MIMLSVM, SISL-Circle, MIML Circle are used for the above purpose. More focus is put on traditional classroom teaching scenes, rather than online courses. MIML circle gave the best result.

New features like the number of days students have been absent are introduced, which led to an increase in accuracy of academic performance prediction by 10-15% compared to when such features were not used [11]. The dataset used contains 500 different records and 17 features. ANN performs better than other classification algorithms used. However, more psychological factors can be taken into account in this research, so as to get a more accurate analysis.

A hybrid classifier is used for predicting students' academic performance. It uses a predictive data-mining model to distinguish between fast and slow learners. Irrelevant inputs were eliminated using sensitivity analysis. It makes use of a hybrid classifier consisting of Fuzzy Artmap and Bayesian Artmap. Performance measures used include MCC, accuracy rate, TP, and FP [12].

The data that was analyzed belonged to students of public schools of brazil. Classification methods based on Gradient Boosting machines are used. 14 nominal and 3 numerical variables were considered. In particular, the variables - 'neighborhood' and 'school' had a major effect on student's performance. Attendance records also played a major role in predicting the performance [13].

The academic performance of students is predicted using several machine learning methods like gradient boosting classifiers and decision trees. Random forest and extreme gradient boosting classification techniques have also been used. [14]. The resultant accuracy achieved of the model is 95% and correlation heatmap is used to study relationships between various attributes. Here, multiclass classification is used instead of binary classification.

Prediction of students' academic performance using Ontological Modelling based on Fuzzy logic. It predicts academic performance in pervasive environments. It takes into account internal and external assessment and viva voce marks for prediction of the final grade. It is implemented in MATLAB [15]. Parameter Tuning can be applied to increase accuracy further.

The use of a new technique was developed by the researchers named MANFIS-S to predict student's performance[16]. MANFIS-S outperformed all other algorithms used for validation which were MANFIS, ANFIS, oneR, Random trees.

Analysis of the factors which influence the students' academic performance prediction using Learning analytics. It uses data from 2 MOOCs on edX on Java programming [17]. The factors which influence this include previous grades, course duration, clickstream data, forum variables, etc.

Both classification and clustering algorithms are employed, and a hybrid model is developed for academic performance prediction. Therefore, techniques used include Support Vector Machine, Decision tree, clustering using K-means, Neural Networks, and Naive Bayes [18]. The results achieved by this hybrid model are better as compared to other models. However, more courses could have been analyzed and more complex algorithms could have been used.

Use of 2 datasets and 5 machine learning algorithms to predict the student's academic performance. Backpropagation and Long-Short Term Memory are used. In addition to this, BP, SVR, and Gradient Boosting Classifiers are used in classification. In conclusion, we see that SVR produced the best result. Even Backpropagation (BP) produced good results [19].

Three classification classifiers- KNN, Naive C4.5, and Bayes are used to create a student academic prediction model. Other factors, along with entrance examination scores like high school awards, admission area, etc. are also considered [20]. The result is that the KNN algorithm is better than C4.5 and Naive Bayes. One limitation could be that it is a binary classifier.

TABLE I. ANALYSIS OF VARIOUS RESEARCH PAPERS

S.No	Authors	Year	Dataset used	Methods used	Result and Remarks
1	V. L. Uskov, J. P. Bakken, A. Shah, A. Byerly	2019	Marks scored by students in assignments of STEM courses.	The Gartner Analytics Ascendancy Model.	Different accuracy- Linear Regression: 96.3%, SVM : 93.78%, RF: 93.41%, Decision tree:92.19%, KNN:69.99%, ANN : 67% and Naive Bayes:55.88% accuracy.
2	W. Nuankaew, J. Thongkam	2020	9500 students of Rajabhat Maha Sarakham University in Thailand	Naive Bayes, SMO, ANN, Random Forest, KNN, Partial decision trees, REPTree	Random Forest gave the best performance, with a precision of 94.70%
3	E. T. Lau, L. Sun, Q. Yang	2019	Chinese University- 1000 students	ANN	Accuracy: ANN -84.8% ,AUC-0.86. One limitation is the poor performance of ANN in classifying students according to gender.
4	R. Ravikumar, V. Akre, F. Aljanahi, A. Rajan,	2018	Dummy grade combinations for students.	Early warning system based on the GPA of students	Students who scored C grade or less than that from the 1st semester have a high chance of falling into risk.
5	Sachio Hirokawa	2018	University Data(480 students)	SVM	Using SVM, accuracy is 84.3% .The author can use more advanced machine learning techniques.
6	H. YAO, D. LIAN, Y. CAO, Y. Wu, T. ZHOU	2019	Data from one university during 2011/09/01 to 2015/06/30	Multi-task learning	The proposed method is very effective but there is a limitation in data collection.
7	S. N. Liao, D. Zingaro, K. Thai, C. Alvarado, W. G. Griswold, L. Porter	2019	Dataset used was obtained from 2 public universities in North America	SVM with radial basis function kernel.	The result improves on the addition of more data. It identifies at-risk students at the earliest point in time, which helps both students and teachers. It also helps in reducing the chance of student failure. The model has not considered courses related to other fields.

8	A. E. Tatar, D. Düstegör	2020	357 students who were admitted to the CCSIT at IAU from Fall 2011 to Fall 2013	Naïve Bayes, logistic regression, random forest	Better accuracy is achieved using the ‘course grade’ variable if the prediction takes place before the third term. Otherwise in all cases using ‘Grade Point Average’ yields the best accuracy. The methodology could be applied on more datasets.
9	E. Admasu, A. Teklay	2019	Dataset was collected from 3 high schools of Ethiopia	Multi-label Ensemble model	SVM performs best in Hamming loss(22.4%), micro F1(83.3%), and macro F1(82.4%), Random Forest performs best in subset accuracy(34.4%), MLP performs best in Jaccard Similarity(72.3%). Partitioning using RAkEL gives better results as compared to a fast-greedy approach.
10	Yuling MA, Chaoran CUI , Jun YU, Jie GUO, Gongping YANG, Y. YIN	2019	Dataset of 1,020 university students from various computer-related courses	SVM, MIMLSVM, SISL-Circle, MIML Circle	MIML-Circle gave the best result. Model needs to be trained with a dataset with more training samples.
11	Sana, Isma Farrah Siddiqui, Qasim Ali Arain	2019	Kiteboard 360 e-learning system(500 records, 17 features)	Naive Bayes classifier, ANN, and Decision Tree	increase in accuracy of academic performance prediction by 10-15% by using new features
12	Roshani Ade	2019	Educational data set	Hybrid classifier consisting of Fuzzy Artmap and Bayesian Artmap.	Hybrid classifier is more accurate. The accuracy was 92% using Hybrid classifier whereas Fuzzy Artmap with 84% accuracy and Bayesian Artmap with 87% accuracy.
13	E. Fernandes, M. Holandaa, M. Victorinoa, V. Borgesa, R. Carvalhoa, G. V. Ervena	2018	School students of Brazil	Classification models that are based on gradient boosting machine	The variables ‘ neighborhood’, ‘school’, and ‘attendance record’ influenced the result greatly.
14	Abhinav Jain, Shano Solanki	2019	UCI	Decision tree, gradient boosting classifier, random forest classifier, and extreme gradient boosting classifier	Random forest classifier Gave the best result. The resultant accuracy achieved of the model is 95%.
15	Ramanathan Veeraiyan, Sivakumar Ramakrishnan	2020	CIA marks, external marks, viva voce marks	Ontological Modeling based on Fuzzy logic	The final grade is directly proportional to the internal assessment and also external assessment whereas viva voce marks slightly influence the final grade.

16	Le Hoang Son · Hamido Fujita	2018	Dataset of students from Vietnam National University and KDD cup datasets as well as the dataset from UCI.	MANFIS-S	MANFIS-S outperformed all the other algorithms used for validation which were MANFIS, ANFIS, OneR, and Random tree.
17	P. M. Moreno-Marcos, T.C. Pong, P. J. M. Merino, C. D. Kloos	2019	MOOC data is collected from edX.	Learning Analytics	A total of 6 models are made. MOOC data is considered. Forum variables were not useful in prediction whereas exercise variables were most useful in prediction. Clickstream data also gave useful predictions in the absence of exercises.
18	B. K. Francis, S. S. Babu	2019	Students' data from Kerala	SVM, Decision tree, K-means, Neural Networks, and Naive Bayes	The hybrid model has an accuracy of 0.7547.
19	B. Sekeroglu, Kamil Dimililer,Kubra Tuncal	2019	Student Performance Dataset and Student's Academic Performance Dataset	Backpropagation, SVR, and Long-Short Term Memory	SVR produced the highest R^2 and EV values
20	Xinyu Du	2019	Students' data from a Chinese University	KNN, Naive C4.5, and Bayes	KNN algorithm is better than C4.5 and Naive Bayes. There is a limitation in data collection due to which attributes may not be optimal

III. FINDINGS OF THE REVIEW

We have studied and analyzed the above research papers and have presented it in a structured format along with the summaries for future references. In our study, we have found out that the variables used for analysis in almost all researches do not include all the relevant factors. Thus, we need a new holistic approach so that we can capture those factors which may not be intuitive but would play a significant role in a student's life. Thus, better surveys are needed to be developed, and new datasets are needed to be generated for a more accurate analysis. We can include factors such as the amount of social media usage, sleeping habits, etc. A thorough analysis of student life is needed.

IV. CONCLUSION AND FUTURE WORK

The purpose of this paper is to find out the methods and techniques which are employed currently for predicting the academic performance of students and also observe the results achieved by them. The machine learning techniques and data mining has been used extensively in this area and provide future scope also to work with huge amounts of data. In the past few decades, numerous studies and researches have already been done on this subject, we have attempted to compare them and put the relevant studies in one place. We conclude that a lot of research can still take place to make academic performance prediction more accurate and globally applicable. Moreover, the practical application of such work is also required and it could be very beneficial for the betterment of our education system.

Many factors that are not usually intuitive but have a direct impact on student's performance need to be considered. More research is also needed in the area of analyzing a student's life from different perspectives and identifying more variables that are left out in the research that we have studied. There also needs to be a system so that an educational institute can directly tap onto the weak points of an individual using the predictions and can give him a more holistic treatment in those specific areas only. There could also be a classification of students based on similar characteristics. A student should also be able to know the areas improving in which would lead to the maximum growth in his performance. So that he could work optimally and more efficiently.

REFERENCES

[1] Sachio Hirokawa. 2018. Key attributes for predicting student academic performance. In Proceedings of the 10th International Conference on Education Technology and Computers (ICETC '18). Association for Computing Machinery, New York, NY, USA, 308–313. DOI:https://doi.org/10.1145/3290511.3290576

[2] W. Nuankaew and J. Thongkam, "Improving Student Academic Performance Prediction Models using Feature Selection," *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Phuket, Thailand, 2020, pp. 392-395, doi: 10.1109/ECTI-CON49241.2020.9158286.

[3] Lau, E.T., Sun, L. & Yang, Q. Modelling, prediction, and classification of student academic performance using artificial neural networks. *SN Appl. Sci.* **1,** 982 (2019). https://doi.org/10.1007/s42452-019-0884-7

[4] R. Ravikumar, F. Aljanahi, A. Rajan, and V. Akre, "Early Alert System for Detection of At-Risk Students,"

2018 Fifth HCT Information Technology Trends (ITT), Dubai, United Arab Emirates, 2018, pp. 138-142, doi: 10.1109/CTIT.2018.8649508.

[5] V. L. Uskov, J. P. Bakken, A. Byerly, and A. Shah, "Machine Learning-based Predictive Analytics of Student Academic Performance in STEM Education," *2019 IEEE Global Engineering Education Conference (EDUCON)*, Dubai, United Arab Emirates, 2019, pp. 1370-1376, doi: 10.1109/EDUCON.2019.8725237.

[6] Huaxiu Yao, Defu Lian, Yi Cao, Yifan Wu, and Tao Zhou. 2019. Predicting Academic Performance for College Students: A Campus Behavior Perspective. ACM Trans. Intell. Syst. Technol. 10, 3, Article 24 (May 2019), 21 pages. DOI:https://doi.org/10.1145/3299087

[7] Soohyun Nam Liao, Daniel Zingaro, Kevin Thai, Christine Alvarado, William G. Griswold, and Leo Porter. 2019. A Robust Machine Learning Technique to Predict Low-performing Students. ACM Trans. Comput. Educ. 19, 3, Article 18 (June 2019), 19 pages. DOI:https://doi.org/10.1145/3277569

[8] Tatar, A.E.; Düştegör, D. Prediction of Academic Performance at Undergraduate Graduation: Course Grades or Grade Point Average? *Appl. Sci.* **2020**, *10*, 4967.

[9] Yekun, E., and Teklay, A., 2020. *Student Performance Prediction With Optimum Multilabel Ensemble Model*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1909.07444> [Accessed 14 September 2020].

[10] Ma, Y., Cui, C., Yu, J., *et al.* Multi-task MIML learning for pre-course student performance prediction. *Front. Comput. Sci.* **14,** 145313 (2020). https://doi.org/10.1007/s11704-019-9062-8

[11] S. B., I. Farrah Siddiqui, y Q. Ali Arain, Analyzing Students' Academic Performance through Educational Data Mining, *3C Tecnología*, pp. 402-421, May 2019.

[12] Ade, R., 2020. *Students Performance Prediction Using Hybrid Classifier Technique In Incremental Learning | International Journal Of Business Intelligence And Data Mining*. [online] Inderscienceonline.com. Available at: <https://www.inderscienceonline.com/doi/pdf/10.1504/IJBIDM.2019.101255> [Accessed 14 September 2020].

[13] Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., and Erven, G., 2020. *Educational Data Mining: Predictive Analysis Of Academic Performance Of Public School Students In The Capital Of Brazil*.

[14] A. Jain and S. Solanki, "An Efficient Approach for Multiclass Student Performance Prediction based upon Machine Learning," 2019 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2019, pp. 1457-1462, doi: 10.1109/ICCES45898.2019.9002038.

[15] Veeraiyan, R., and Ramakrishnan, S., 2020. [online] Ijaema.com. Available at: <http://ijaema.com/gallery/271-january-3311.pdf> [Accessed 14 September 2020].

[16] Son, L.H., Fujita, H. Neural-fuzzy with representative sets for prediction of student performance. *Appl Intell* **49,** 172–187 (2019). https://doi.org/10.1007/s10489-018-1262-7

[17] P. M. Moreno-Marcos, T. Pong, P. J. Muñoz-Merino, and C. Delgado Kloos, "Analysis of the Factors Influencing Learners' Performance Prediction With Learning Analytics," in IEEE Access, vol. 8, pp. 5264-5282, 2020, doi: 10.1109/ACCESS.2019.2963503.

[18] Francis, B.K., Babu, S.S. Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *J Med Syst* **43,** 162 (2019). https://doi.org/10.1007/s10916-019-1295-4

[19] Boran Sekeroglu, Kamil Dimililer, and Kubra Tuncal. 2019. Student Performance Prediction and Classification Using Machine Learning Algorithms. In <i>Proceedings of the 2019 8th International Conference on Educational and Information Technology</i> (<i>ICEIT 2019</i>). Association for Computing Machinery, New York, NY, USA, 7–11. DOI:https://doi.org/10.1145/3318396.3318419

[20] Du, X. (2019). A Prediction Model for High School Students' Academic Performance in College Based on Machine Learning.