A Survey of Different Supervised Learning-Based Classification Models for Student's Academic Performance Prediction

Sandeep Kumar and Ritu Sachdeva

Abstract Despite delivering high-quality learning, the need to evaluate student's academic achievement has become increasingly essential to optimize the integrity and aid learners to achieve excellent results in academics. One of the critical challenges is the inadequacy of an accurate and efficient estimation method. Predictive analytics (PA) can help organizations make more intuitive and intelligent decisions. The purpose of this paper is to evaluate existing educational-based student performance analytics study that focuses on forecasting learner educational excellence. Earlier academics have presented several strategies for developing the optimal process framework, employing various academic statistics, methodologies, methods, and platforms. Numerous learning challenges, like categorization, prediction, and cluster analysis, are associated with the predictive Analysis used during estimating students' achievement. The student performance prediction model has various advantages and applications, such as it is used to help instructors in curriculum design and improvement. SPP provides recommendations to the students and offers comments to educators. In this paper, several methods of student performance prediction (SPP) are compared with the help of different performance parameters such as accuracy, specificity, and sensitivity.

Keywords Student performance · Prediction models · Predictive analytics · Academics-based learning models

S. Kumar (✉) · R. Sachdeva
Department of Computer Science and Engineering (CSE), Lingaya'sVidyapeeth, Faridabad, Haryana, India
e-mail: Katariasandeep90@gmail.com

R. Sachdeva
e-mail: dr.ritu@lingayasvidyapeeth.edu.in

D. Gupta et al. (eds.), *International Conference on Innovative Computing and Communications*, Lecture Notes in Networks and Systems 492,
https://doi.org/10.1007/978-981-19-3679-1_44

1 Introduction

Educational data mining is a relatively new study subject that attempts to understand hidden connections in various educational scenarios, such as learner data analysis, student learning activity recognition, instructor course design, and academic planning, and scheduling. Student academic performance is defined from different perspectives, and quantifiable assessment serves an essential part in today's educational institutions. Student performance prediction (SPP) certainly makes sense [1]. SPP could support learners in selecting appropriate programs or activities and making educational strategies. SPP can assist educators in modifying educational content and teaching approaches related to student capability and identifying at-risk individuals. SPP can assist educational administrators in evaluating the curriculum program and improving the coursework. Generally, educational development participants could produce better initiatives to expand academic attainment.

Furthermore, the data-driven SPP review attempts as an unbiased benchmark for the education sector. In diverse contexts, student performance prediction (SPP) can be expressed as various issues. The most critical process of data mining in education is predicting the students' performance. It examines online information and uses several approaches and models such as correlation analysis, neural network models, rule-based frameworks, regression, and Bayesian networks [2]. Based on characteristics retrieved from information filtering, such an approach allows everyone to anticipate the performance of the students, i.e., forecast his performance in a program and their performance rating. The classification and prediction techniques help detect undesired academic achievements such as incorrect activities, decreased morale, cheating, and underachievement. Segmentation, grouping, anomaly analysis, feature engineering, logistic regression, and artificial neural are the most commonly utilized student performance prediction (SPP) algorithms.

The section categorization of this paper is as follows: In Sect. 2, several existing techniques of student performance prediction (SPP) are surveyed. Sect. 3 discussed academic prediction analysis's applications and advantages, and Sect. 4 described several student prediction models with existing performance evaluations with different classification models. Sect. 5 elaborates the theoretical analysis conclusion of the several learning models.

2 Literature Review

Several existing methods of student performance prediction models are reviewed in this section.

Chango et al. [3] designed a data fusion approach based on blended learning. The information fusion system was used to determine university students' overall educational excellence integrating different, multidimensional data across blended educational contexts. Information about first-year university graduates was collected

and pre-processed from a diversity of ways, including classroom sessions, practical classes, interactive moodle workshops, and a midterm test. The main goal was to determine where the information fusion method yields the most extraordinary outcomes with our dataset. The findings indicate that aggregates and picking the best characteristics method with fractional order data generate the best predictions. Silva et al. [4] proposed an artificial intelligence model to predict academic performance. The neural network was used for the analysis of the performance of students. A backward propagating approach was used to develop a multilayer perceptron neural network to classify the chance to dominate the competition. The classification performance rates were very high, with an average of 74.98%, including all programs, demonstrating the determinants' usefulness in forecasting educational attainment. Two estimation techniques, specifically student evaluation ratings and ultimate performance of students, were developed by Alshabandar et al. [5]. The algorithms could identify the determinants that affect MOOC students' educational objectives. Mainly as a consequence, all methods perform practical and precise measurements. The most negligible RSME improvement was produced by RF, with an overall average of 8.131 general students' evaluations grading system.

In contrast, GBM yielded the best performance in the ultimate version of the student, with an overall average of 0.086. Al Nagi et al. [6] used multiple classification algorithms on an educational database for courses online to select the optimum model to classify academic achievement based on critical variables that may lead to different results. Artificial neural network (ANN), decision tree (DT), KNN, and support vector machine (SVM) were some of the classifiers employed. Experiments were performed with actual statistics, as well as the algorithms were assessed using four performance indicators: precision, accuracy, f-measure, and recall. Raga et al [7] used the deep neural network design and Internet communications features as training sets. The authors were investigated with constructing a forecasting model for academic success in the initial stages of the teaching and learning process. Firstly, several measurements were carried out to find the model parameters for just the highest Convolutional Neural Network (CNN) architecture, as it was used as a foundation classification model. This result's accuracy for forecasting exam findings for a specific course was 91.07% with a ROC and AUC value of 0.88.

In contrast, the overall efficiency with forecasting midterms consequences was 80.36%, with a ROC AUC value of 0.70. Czibula et al. [8] presented Students performance prediction using relational association rules (S-PRAR). This unique categorization approach relies on interpersonal sequential pattern development to forecast an academic program's outcome regarding student ratings during the academic session. Investigations on three multiple datasets acquired from Romania's Babes-Bolyai University revealed that the S-PRAR was implemented to solve well. The S-PRAR classification algorithm presented in the proposed research benefited from becoming generalized since it was not limited to the learners' achievement classification step. Table 1 discussed various existing student academic performance prediction models depicted with research gaps, comparative techniques, future work, and performance metrics.

Table 1 Comparative analysis of various existing methods of student performance prediction models

Author name	Proposed methods	Research gaps	Performance metrics	Dataset	Future scope
Chango et al. [3]	Data diffusion-based approach	Need to extract semantic level features	Accuracy, The area under the roc curve	Data collected from UCO (University of Cordoba), Spain	Knowledge-based Fusion technique will be implemented for more effective results
Silva et al. [4]	Multi perceptron neural network	Poor classification	Accuracy	Data gathered from industrial engineering race university	Prediction accuracy will be improved with a hybrid technique
Alshabandar et al. [5]	Machine learning-based approach for performance prediction	Issues in features selection. And need to work on temporal based features	Accuracy, F1-score, Sensitivity, Specificity, AUC	Open university learning analytics dataset	Temporal features will be used for effective outcomes
Al Nagi et al. [6]	SVM, ANN, Decision tree, KNN	Poor feature extraction results	F-measure, Accuracy, Recall, Precision	Open university learning analytics dataset	The extraction technique will be enhanced for better feature extraction
Raga et al. [7]	The deep neural network-based system	Limited information	Accuracy, AUC	MAT and COM datasets	A pre-trained feature selector will be used for effective outcomes
Czibula et al. [8]	Relational association rule mining categorization technique	Fail to solve regression issues Limited to solve only binary issues	Recall, sensitivity, specificity, *F*-score, AUC (Area under the ROC curve)	Data collected from Babes-Bolyai University	Gradual relational association rules will be implemented for the efficient outcomes

3 Applications and Advantages of Student Performance Models

3.1 Applications

There are various applications of the student prediction performance [9, 10], and some of the applications are (i) Assessing students for enhancing their outcomes*:* The main objective of the student performance prediction model is to monitor the performance of the students and help the learners to improve their performance. (ii) Help instructors in curriculum/course design and improvement*:* Student performance prediction (SPP) helps the instructors to design the program course. Also, it helps to analyze students' interests. Students can quickly learn the intersecting content; therefore, SPP gives the instructors directions for designing the intersecting course. (iii) Providing comments to educators*:* Students' performance can be predicted with the help of data mining approaches. The approaches are used to analyze the student's achievements and based on achievement analysis. Comments are provided to educators. (iv) Student's recommendations: Recommendation systems can also be used to tap into scientific relations collected in course-learning databases. This recommender platform's objective is to facilitate learners across a whole program using a competency-based evaluation approach. Learners must attain escalating levels of achievement with each program competency through productive projects. Learners may find suggestions helpful in advancing toward the next level of expertise.

3.2 Advantages

The applications mentioned earlier of student prediction performance provided various advantages: Improves student's self-reflection: The student performance prediction model stores the overall information of the students, which can be used to predict the academic performance [2]. Whenever students lack in any course, it provides an alarm to students. Also, it gives a performance chart of the students that can reflect the students learning. Identifying Unwanted Student Behaviors: The student performance prediction model helps determine students' erratic and unwanted behavior. Student's poor performance risk identification: Proactive advising of the student performance prediction model can aid students in a reasonable timeframe using the information to identify vulnerable learners in a program. Academic advisers and student achievement professionals need meaningful information regarding learner performance and outcomes, which the student performance prediction model provides. Measure the impact of tool adoption: Universities may accurately assess how and why the platform is being implemented through access to the information on improving the educational activities and promoting the implementation where it will have the most impact. Universities can also evaluate third-party

services' usage; a small number of academic staff may employ that and therefore do not reflect the hefty premium.

4 Supervised-Based Learning Classification Models for Student Performance Prediction

Predictive modeling is commonly used within educational data mining methods to determine academic achievement. Numerous activities are employed to develop predictive modeling, including categorization, Analysis, and segmentation. Categorization is the most common requirement used to determine academic achievement. Under the categorization issue, numerous techniques have been used to estimate academic achievement [11].

4.1 Artificial Neural Networks (ANN)

An artificial neural network combines multiple nodes and works like a human brain system. An artificial neural network is another commonly used technique for student performance prediction. The advantage of ANN is that it can find the relation between the large dataset and search for every possible response. ANN is a group of interlinked input/outcome labels, and a load exists on each link**.** At the time of the training stage, the system acquires knowledge through load arrangement to estimate the accurate labels of the input module [12]. It is exceptionally capable of deriving explanations from complex or non-specific data. The basic structure of the artificial neural network is presented in Fig. 1.

4.2 Convolution Neural Network (CNN)

CNN is a deep learning technique for processing information. A convolutional neural network is a branch of artificial intelligence that collects information using convolutional layers, a computer vision unit method. It is designed as hierarchical spatial features from lower to higher-level patterns [13]. It is comprised of pooling, convolution, and fully connected layers. When the layers are weighted, then the structural design is created. The main functions of the layers of convolution neural network are categorized into different areas: The input layer may store the pixel values of the picture. The basic architecture of the convolutional neural network is presented in Fig. 2.

The convolution layer may identify the output value of neurons that connect to the local area of the input by calculating the scalar product among the weights and area

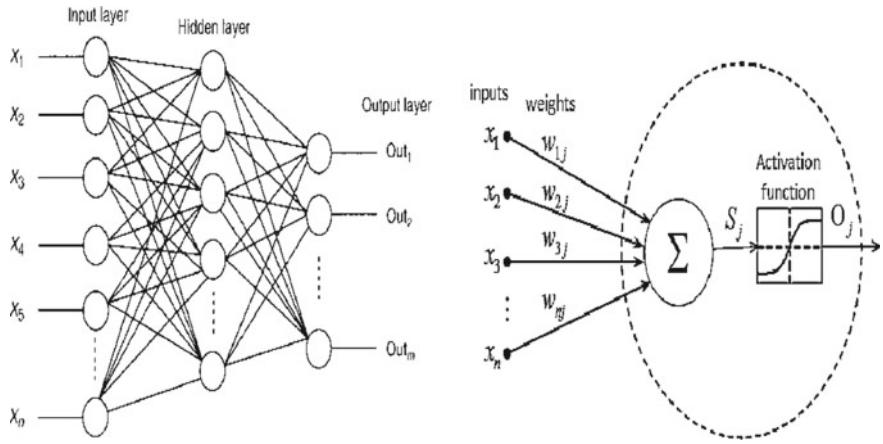

Fig. 1 The basic structure of ANN model [12]

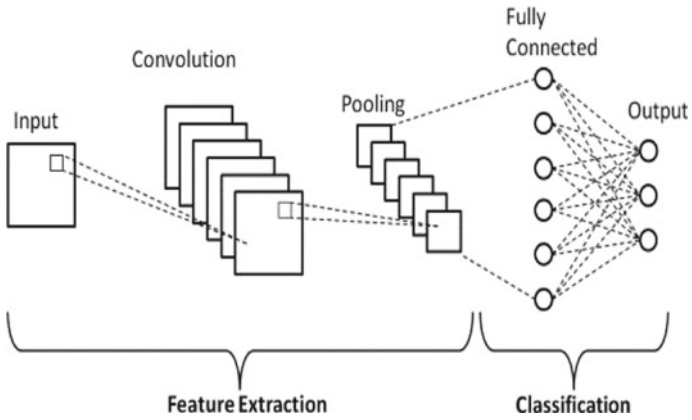

Fig. 2 The basic architecture of CNN model [13]

attached to the information. Thus, the rectified linear unit aimed to put on activated value like sigmoid to result of activation generated by the last layer. The pooling layer may execute the down sample with the spatial dimension of required input and decrease the number of metrics with the activated value. And, fully connected layer performs a similar function searched in standard neural network and creates the class scoring from the activators for the classification purpose.

4.3 Support Vector Machine

The (support vector machine (SVM) was created with binary classification in consideration. Several categorization methods have been proposed to generalize SVM to the multi-class case. SVM is the simplest way to classify binary data classes [14]. SVM, one of the machine learning algorithms based on supervised learning, helps in classification and regression. The main objective of SVM is to provide a hyperplane and generate the different class clusters [15].

4.4 Decision Tree

The decision tree is a prediction and classification technique. The structure of the decision tree is a flowchart-like structure. Each node is connected to the other and forms a tree structure. The internal node denotes test attributes; the test outcome is presented with branches [16]. The class label is shown as leaf nodes or terminal nodes. The trees inside the decision tree framework can be trained by dividing the source set into small subsets based on the test attribute values.

4.5 K-NN Model

The *K*-NN technique ensures that the particular incoming instance and existing cases are equivalent and assigns the new case mostly in subcategories that are most compatible with the existing subcategories [17]. The *K*-NN method accumulates all available information and identifies a subsequent set of statistics premised on its resemblance to the current data. It implies that new information can be quickly sorted into a well-defined subcategory that uses the *K*-nearest neighbor method. The *K*-NN approach is used for regression and classification issues, but it is more generally applied for classification methods.

4.6 Random Forest

Random forest is a flexible, straightforward computational model in the vast majority of circumstances, produces tremendous success with hyper-parameters or without hyper-parameters modification. Along with its simplicity and versatility, it has become one of the most commonly used approaches. It can be used for classification and regression tasks. The essential characteristics of the RF algorithm are that it can manage sets of data with both categorical and continuous, as in regression and classification issues [17].

4.7 Fuzzy Logic

Fuzzy logic is a method of variables computing that enables the computation of numerous possible conditional probabilities using the exact attributes. Fuzzy logic is a type of logic that is used to simulate human understanding as well as thinking. It is a computing technique that focuses on "truth degree" instead of the conventional "correct or incorrect" (1 or 0) binary decision logic that the digital machine is built on "fuzzy" refers to something unclear or ambiguous [18]. A fuzzy system can be described as a set of IF–THEN logic containing fuzzy propositions, a mathematical or differential calculation containing random variables as variables that represent the uncertainty of attribute values. In Table 2, various existing methods of student performance prediction (SPP) with attributes and performance metrics.

The comparison of various existing methods of student performance prediction (SPP) is presented in Figs. 4 and 5. The comparison analysis of different current methods of student performance prediction provided that the naïve Bayes method has attained maximum accuracy, precision, and recall compared to the other academic prediction models.

Table 2 Existing methods of student performance prediction (SPP)

Methods	Attributes	Recall	Accuracy	Precision
ANN [6]	Evaluation score	0.89	0.56	0.57
CNN [13]	Attendance	–	0.76	–
SVM [13]	CGPA	0.65	0.6488	0.64
The decision tree [16]	CGPA	–	0.85	–
K-NN [17]	Gender	62.9	0.63	63.4
Naïve Bayes [17]	Gender	93.6	0.93	93.17

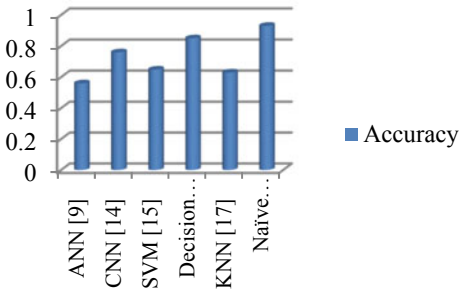

Fig. 4 Comparison between several existing models with an accuracy rate

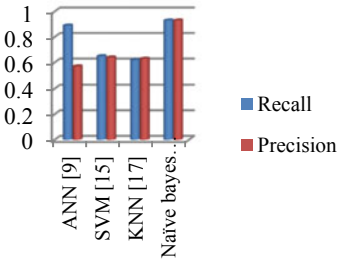

Fig. 5 Comparison between several existing models with recall and precision rate

5 Conclusion and Future Work

This paper analyzed recent research on student performance prediction (SPP) models. Estimating student achievement is the most efficient approach for learners and educators to enhance learning by ensuring that students complete their courses on time. Existing research has used a range of methodologies to develop the best forecasting models. Many factors were selected and evaluated to identify the most critical and robust characteristics to estimate an optimal framework. Attendance, CGPA, race, and evaluation score have all been used by the majority of the investigators. Several features have an immense impact on whether or not a learner will complete their studies. Among the most critical aspects of evaluating academic achievement are the forecasting techniques. Several researchers use the decision tree, artificial neural network, support vector machine (SVM), *K*-Nearest Neighbor *K*-NN, and other categorization, prediction, and segmentation techniques. Many use a combination of strategies to create a more reliable system with higher predicted accuracy. The comparison analysis of various existing methods of student performance prediction provided that the naïve Bayes method has attained maximum accuracy, precision, and recall. Further work will introduce the novel feature extraction and prediction models with mathematical expressions to fetch the reliable feature values and improve the classification metrics.

References

1. Zhang Y, Yun Y, An R, Cui J, Dai H, Shang X (2021) Educational data mining techniques for student performance prediction: method review and comparison analysis. Front Psychol 12
2. Agrawal H, Mavani H (2015) Student performance prediction using machine learning. Int J Eng Res Technol 4(03):111–113
3. Chango W, Cerezo R, Romero C (2021) Multi-source and multimodal data fusion for predicting academic performance inblended learning university courses. Comput Electr Eng 89:106908
4. Silva J, Romero L, Solano D, Fernandez C, Lezama OBP, Rojas K (2021) Model for predicting academic performance through artificial intelligence. In: Computational methods and data engineering, Springer, Singapore, pp 519–525
5. Alshabandar R, Hussain A, Keight R, Khan W (2020, July) Students performance prediction in online courses using machine learning algorithms. In: 2020 International joint conference on neural networks (IJCNN), IEEE, pp 1–7

6. Al Nagi E, Al-Madi N (2020, October) Predicting students performance in online courses using classification techniques. In: 2020 International conference on intelligent data science technologies and applications (IDSTA), IEEE, pp 51–58
7. Raga RC, Raga JD (2019, July) Early prediction of student performance in blended learning courses using deep neural networks. In: 2019 International symposium on educational technology (ISET), IEEE, pp 39–43
8. Czibula G, Mihai A, Crivei LM (2019) S PRAR: a novel relational association rule mining classification model applied for academic performance prediction. Proc Comput Sci 159:20–29
9. Kabakchieva D (2012) Student performance prediction by using data mining classification algorithms. Int J Comput Sci Manage Res 1(4):686–690
10. Jacob J, Jha K, Kotak P, Puthran S (2015, October) Educational data mining techniques and their applications. In: 2015 International conference on green computing and internet of things (ICGCIoT), IEEE, pp 1344–1348
11. Shahiri AM, Husain W (2015) A review on predicting student's performance using data mining techniques. Proc Comput Sci 72:414–422
12. Zacharis NZ (2016) Predicting student academic performance in blended learning using artificial neural networks. Int J Artif Intell Appl 7(5):17–29
13. Ma Y, Zong J, Cui C, Zhang C, Yang Q, Yin Y (2020, January) Dual path convolutional neural network for student performance prediction. In: International conference on web information systems engineering, Springer, Cham, pp 133–146
14. Burman I, Som S (2019, February) Predicting students academic performance using support vector machine. In: 2019 Amity international conference on artificial intelligence (AICAI), IEEE, pp 756–759
15. Support vector machine (SVM) algorithm-Javatpoint, 2021. www.javatpoint.com. [Online]. Available: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm. Accessed 24 Dec 2021
16. Pandey M, Sharma VK (2013) A decision tree algorithm is pertaining to the student performance analysis and prediction. Int J Comput Appl 61(13)
17. Amra IAA, Maghari AY (2017, May) Students performance prediction using KNN and Naïve Bayesian. In: 2017 8th International conference on information technology (ICIT), IEEE, pp 909–913
18. Barlybayev A, Sharipbay A, Ulyukova G, Sabyrov T, Kuzenbayev B (2016) Student's performance evaluation by fuzzylogic. Proc Comput Sci 102:98–105