International Journal of Computer Sciences and Engineering
Vol.12, Issue.9, pp.09-13, September 2024
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

Analysis of Students Performance Prediction Models Using Machine Learning Approaches

D. Boominath[1*], S. Dhinakaran[2]

[1,2]Dept. of Computer Science, Rathinam College of Arts & Science, Coimbatore, India

**Corresponding Author: boominath4u@gmail.com*

Received: 18/Jul/2024; **Accepted**: 20/Aug/2024; **Published**: 30/Sept/2024. **DOI**: https://doi.org/10.26438/ijcse/v12i9.913

Abstract: The field of Educational Data Mining (EDM) is still young and is focused on improving existing data mining (DM) techniques as well as creating new ones for locating data originating from educational systems. It seeks to employ these techniques to arrive at a logical understanding of students and the kind of learning environment they ought to experience. Knowledge Tracing (KT) and the prediction of student performance are closely intertwined. The academic community has made an effort to address it and has produced findings that are competitive. Several strategies have been implemented over the past 20 years that have improved on already-existing techniques by attacking the issue from different model architectures and experimenting with various datasets and formats. The efficiency of various machine learning models for predicting student performance is examined in this research. The outcomes were contrasted with earlier research that forecasted student achievement.

Keywords: Educational Data Mining, Machine Learning, Student Performance Prediction, Knowledge Tracing, and Classification.

1. Introduction

One of the study fields is called Educational Data Mining (EDM), which applies data mining techniques on educational datasets to extract important knowledge and make it comprehensible and interpretable for decision-making. There are a number of variables that influence the decision to forecast and track student performance in educational institutions. The EDM has a series of algorithms that mine the academic records of the students for hidden patterns. The two primary categories of algorithms are supervised learning techniques and unsupervised learning techniques. Utilizing labeled training data, supervised learning techniques are utilized to evaluate a classifier and predict the label of a future or unlabeled record being tested. The goal of educational data mining is to create, investigate, and use computerized methods to find patterns in huge amounts of educational data that would otherwise be difficult or impossible to evaluate due to the vast amount of data they exist within.

In educational data mining, there are numerous well-liked techniques. Prediction, classification, grouping, and regression are some of the commonly utilized ones. In order to forecast performance or learning outcomes using the data from the existing data, prediction tries to establish patterns. Predicting a student's course activity is the goal. The methods most frequently employed for achieving this type of objective include classification, clustering, and association. Prediction has been used in educational data mining to forecast student achievement and identify student behaviour. Prediction attempts to deduce a specific characteristic or certain element of the data from other aspects.

Depending on the type of labels in the training dataset, supervised learning uses either classification or regression approaches. Unsupervised approaches use inherited correlations between records to extract hidden knowledge from unlabeled datasets, with the results serving as a label for subsequent records. The clustering technique, which divides data into groups or clusters based on their associated qualities, is the most popular unsupervised method.

The most widely used methods to address the issue of predicting student performance are classification techniques. For the purpose of creating efficient classifiers, these strategies rely on training sets of features taken from history records of students as well as students' actual performance. In order to forecast the performance of the current students, the classifiers represent inherited patterns extracted from the training sets of previous students' records. The training phase and the testing phase are the two key stages used by classification algorithms to build a classifier model. A training set is utilized by the classification approach to create a classification model during the training phase, and a test set is used during the testing phase to assess the classifier model.

The main goal of this study is to identify underachievers early in their academic careers in order to retain as many students as possible and ensure that each student achieves academic success. Early identification of these pupils will aid in developing a plan to give them extra attention to assist them improve their academic performance. This research will assist in enhancing learning processes, teaching strategies, and educational quality. It can also aid in decisions about selecting teachers and enhancing student performance, which will enhance educational outcomes generally.

The remaining sections of the paper are organised as follows. Section 2 highlights previous studies in the prediction and classification of student performance. Section 3 explains the problem statement. Section 4 examines the comparison study of the various machine learning model for students' performance prediction, and Section 5 that gives the proposed work's conclusion.

2. Literature Survey

The primary focus of many researches is establishing a prediction model as well as elements that influence a student's academic performance. a thorough evaluation of the research on data mining strategies for forecasting student success. Their paper's major objective was to give a summary of the data mining methods that have been applied to forecast students' academic achievement and how the prediction algorithm may be used to pinpoint the most crucial characteristics in a student's data. To develop a prediction model for student academic performance for certain courses or disciplines, researchers have carried out a number of investigations. These studies use a number of student data types and characteristics to identify and categorize their participants.

While attempting to clarify diverse professional viewpoints on EDM and its technique, this chapter also covers many studies that used EDM to examine educational data, in particular higher educational data, in order to identify relationships between various parameters and the performance of HE students.

According to Balaji, P., et al. (2021), machine learning is becoming more and more significant as a tool for decision assistance across a wide range of study fields. The authors argue that because each work submitted for review uses a different dataset and lacks benchmark datasets, it would not be possible to identify a specific machine learning model for the purpose of predicting student academic achievement. However, the use of machine learning techniques in educational mining is currently limited, and more research should be done in order to produce outcomes that are well-formed and generalizable.

According to C. F. Rodrguez-Hernández et al.'s analysis published in 2021, artificial intelligence applications in education have grown recently. However, to further the systematic application of these approaches, more conceptual and methodological knowledge is required. Analysis of the significance of a number of well-known predictors of academic performance in higher education was the main goal. 162,030 students from Colombian private and state colleges, including both genders, were included in the sample. The results indicate that artificial neural networks can be used to accurately classify students' academic achievement as either high (accuracy of 82%) or low (accuracy of 71%). Other machine-learning methods are outperformed by artificial neural networks in evaluation measures like recall and the F1 score.

A lack of balanced data processing techniques that can effectively capture student attributes and accomplishment was addressed by Wang, X., et al. in 2023. The authors realized that predicting students' academic achievement was turning into a crucial service in a system of intelligent education backed by computers. The suggested consists of three parts: a module for collaborative data processing to improve data quality, a module for scalable metadata clustering to balance out the imbalance of academic features, and a module for XGBoost-enhanced SAP prediction to forecast academic performance. With almost 98% of actual samples, the findings lead to an improved out-sample fit that reduces prediction errors between 1% and 9%..

According to Aslam, N., et al. (2021), Deep Learning (DL) models led to accurate data prediction. The authors employed the DL model to forecast student performance. SMOTE (synthetic minority oversampling technique) is used to address the imbalance problem in the data set. All feature sets, with the exception of G2 and G3, are used to evaluate the performance model, along with evaluation metrics like precision, recall, F-score, and accuracy. The outcomes demonstrated the value of the suggested DL model in making early predictions about the academic achievement of the pupils. The model's accuracy was 0.964 for the data set from the Portuguese course and 0.932 for the data set from the mathematics course.

Ensemble Methods were suggested by Ajibade, S. S. M., et al. in 2022 to increase the precision of student performance predictions. In e-learning systems or web-based education, student behavioral traits are crucial since they show how engaged the student is with the system. The included dataset was subjected to feature analysis, followed by data preprocessing, which is an essential stage in the knowledge discovery process. The preprocessed dataset is categorized using classifiers like Nave Bayes (NB), Decision Tree (ID3), Support vector machines (SVM), and K-Nearest Neighbor (KNN) in order to predict student academic achievement. The proposed model's accuracy is increased by the use of ensemble methods. Common ensemble techniques that we used include bagging, boosting, and voting algorithms. Using ensemble techniques, we were able to achieve a better result, demonstrating the proposed model's reliability.

Numerous research on student behavior have been done in order to predict student success. For this task, students were categorized based on their learning activity using a variety of data mining approaches. Classification is the process of organizing data into predetermined groups and classes. It is also referred to as supervised learning and incorporates both learning and classification. During the learning phase, a classification algorithm analyzes training data, and during the classification phase, test data is utilized to gauge the precision of the classification rules.

3. Problem Statement

In the field of education, the development of intelligent technologies is gaining traction. The rapid growth of educational data indicates that traditional processing methods may be limited and inadequate. Reconstructing data mining research technique has therefore gained prominence in the sphere of education. in order to monitor the pupils' upcoming performance and prevent irrational evaluation findings. EDM gives teachers the ability to analyze internal elements affecting students' performance, create statistical methodologies to anticipate students' academic achievement, and predict circumstances like dropping out of school or losing interest in the course. In order to forecast student performance, pinpoint slow learners, and identify dropouts, numerous DM techniques are used. Early prediction is a recent phenomenon that uses evaluation techniques to help pupils by suggesting relevant remedial measures and laws in this field.

Predicting and identifying students' academic performance in order to mentor them toward improved performance and deliver quality education. Higher education institutions' principal goal is to give their pupils a high-quality education. To identify the low performers early on, a reliable performance prediction of the students is helpful. The goal of this research is to identify and collect information for predicting both good and bad performances.

Table 1

S. No.	Name of the Author(s) & Year	Proposed Methodology	Merit(s) / Demerit(s)
1.	Kaunang, F. J., & Rotikan, R. (2018	Classification and Regression Trees (CART)	**Merit(s): -** ➢ Data manufacturing requires less endeavor throughout preprocessing compared to sordid algorithm results trees. ➢ Data values slave no longer affect the technique over building a selection arbor to somebody tremendous extent. **Demerit(s): -** ➢ A decision tree is difficult compared to sordid algorithms because of prediction. ➢ Applying the decision tree algorithm and continuous values lag is much less predictable.
2.	Hasan, R., et al., (2018)	A Linear Model Hybrid Random Forest (HRFLM).	**Merit: -** ➢ It usage a rule-based strategy as a substitute for scale estimate. No function scaling (standardization or normalization) is required within the case on Random Forest. **Demerit: -** ➢ Random Forest takes an extra brush time to educate compared to the choice bushes up to expectation to fulfill a bunch concerning trees.
3.	Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015)	SVM, ANN and Decision Tree.	**Merit(s): -** ➢ SVM event is an outskirt on dissolution in classes. ➢ Synthetic neural networks need numerical electricity, which may be employed with some identical times. **Demerit(s): -** ➢ SVM, stop tree, and then ANN truth were 83.4%, 83.3%, and 80.1%. ➢ It may also not work well because we hold even a great deal, and coaching day is required all through the massive data set.
4.	Imran, M., et al. (2019)	Recurrent Neural Network (RNN)	**Merit(s): -** ➢ RNN inputs execute use inner devotion for the non-existent function to condition the free series litigation to enter, and the mannequin quantity does not increase. **Demerit: -** ➢ By its repetitive nature, the score is slow. RNN fashions can lie hard following the train.
5.	Ha, D. T., et al. (2020)	Fuzzy C-Mean (FCM)	**Merit: -** ➢ An ambiguous C algorithm (FCM) is a law that approves a piece of data after belonging to a couple or more clusters. **Demerit(s): -** ➢ Long computation time. ➢ Sensitivity to noise yet membership dimensions for a predicted low (or no) acquirements (noise points).

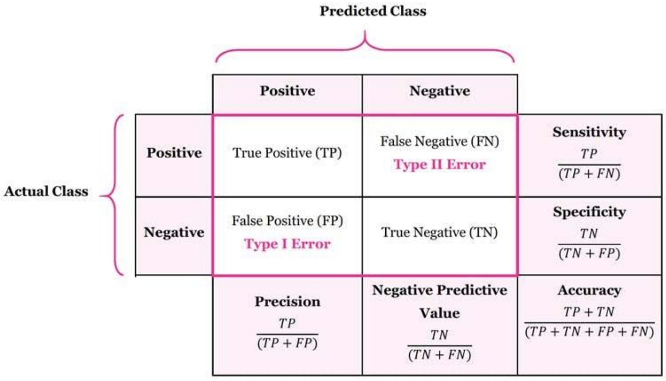
4. Review of Machine Learning Models

4.1. Performance Measures

4.1.1. Accuracy

One of the most widely utilised measures for rating performance is accuracy [15]. It calculates the fraction of correctly categorised student answers by dividing the number of correctly categorised answers by the total number of answers.

Accuracy = TruePositives / (TruePositives + FalsePositives)



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 1

5. Conclusion

The amount of information kept in a database for education at IHL is growing quickly over time. The classification techniques are applied to the student data in order to extract information about the student from such a vast amount of data and to identify the factor that contributed to the students' performance. The majority of publications simply addressed one component of accuracy, and it appeared to be a biased one. In fact, a wide range of metrics that are appropriate for the study's problem can be used to determine the performance measures, such as classification or regression. It is typical practice when a model is proposed to compare how different ML models perform on the data that has been gathered, which may affect how accurate or reliable the data is. However, it is best practice to compare the proposed model's performance to datasets used in previous research studies in order to demonstrate the model's accuracy. This will likely result in the model being fine-tuned to fit more datasets.

Conflict of Interest

The authors affirm that there are no conflicts of interest with any entities or individuals concerning the subject matter discussed in this paper. No financial or non-financial support has been received from any parties associated with the content of this review. Our conclusions and viewpoints are based on an impartial and objective assessment of the existing research and data.

Funding Source

This study was conducted independently without specific grant support from public, commercial, or non-profit funding organizations. It was carried out without external financial assistance, ensuring the neutrality and objectivity of the research.

Authors' Contributions

Both authors made significant contributions to this research. Author 2 was responsible for the literature survey and problem statement. Author 2 managed and review of machine learning models and the conclusion. All authors participated in the development, review, and editing of the manuscript.

Acknowledgements

The completion of this paper and the underlying research would not have been achievable without the support of the Management, CEO and Principal of Rathinam College of Arts and Science.

References

[1] Kaunang, F. J., & Rotikan, R, "Students' academic performance prediction using data mining". In the proceedings of the 2018 Third International Conference on Informatics and Computing (ICIC) IEEE, pp.**1-5, 2018.**

[2] Hasan, R., Palaniappan, S., Raziff, A. R. A., Mahmood, S., Sarker, K. U, "Student academic performance prediction by using decision tree algorithm". In the proceedings of the 2018 4th international conference on computer and information sciences (ICCOINS) IEEE, pp.**1-5, 2018.**

[3] Ahmad, F., Ismail, N. H., Aziz, A. A, "The prediction of students' academic performance using classification data mining techniques", Applied mathematical sciences, Vol.**9**, Issue.**129, 2018.**

[4] Balaji, P., Alelyani, S., Qahmash, A., Mohana, M, "Contributions of machine learning models towards student academic performance prediction: a systematic review", Applied Sciences, Vol.**11**, Issue.**21, 2015.**

[5] Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., Cascallar, E, "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation", Computers and Education: Artificial Intelligence, Issue.**2**, pp.**100018, 2021.**

[6] Wang, X., Zhao, Y., Li, C., Ren, P, "ProbSAP: A comprehensive and high-performance system for student academic performance prediction", Pattern Recognition, Vol.**1**, Issue.**7**, **2023.**

[7] Aslam, N., Khan, I., Alamri, L., Almuslim, R, "An improved early student's academic performance prediction using deep learning", International Journal of Emerging Technologies in Learning (iJET), Vol.**16**, Issue.**12**, pp.**108-122, 2021.**

[8] Ajibade, S. S. M., Dayupay, J., Ngo-Hoang, D. L., Oyebode, O. J., Sasan, J. M. "Utilization of Ensemble Techniques for Prediction of the Academic Performance of Students", Journal of Optoelectronics Laser, Vol.**41**, Issue.**6**, pp.**48-54, 2022.**

[9] Imran, M., Latif, S., Mehmood, D., Shah, M. S, "Student academic performance prediction using supervised learning techniques". International Journal of Emerging Technologies in Learning, Vol.**14**, Issue.**14, 2019.**

[10] Ha, D. T., Loan, P. T. T., Giap, C. N., Huong, N. T. L, "An empirical study for student academic performance prediction using machine learning techniques", International Journal of Computer Science and Information Security (IJCSIS), Vol.**18**, Issue.**3**, pp.**75-82, 2020.**

[11] Lu, O. H., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., Yang, S. J, "Applying learning analytics for the early prediction of Students' academic performance in blended learning", Journal of Educational Technology & Society, Vol.**21**, Issue.**2**, pp.**220-232. 2018.**

[12] Asiah, M., Zulkarnaen, K. N., Safaai, D., Hafzan, M. Y. N. N., Saberi, M. M., Syuhaida, S. S, "A review on predictive modeling technique for student academic performance monitoring", In the proceeding of the MATEC Web of Conferences, EDP Sciences, Vol.**255**, pp.**03004, 2023.**

[13] Balaji D. Raj, Marimuthu A, "Classification of ECG Signals Using Self Advising Support Vector Machine and Fuzzy C Means Clustering", International Journal of Research in Advent Technology, Vol.**6**, No.**11**, pp.**3252-3259, 2018.**

[14] Walia, N., Kumar, M., Nayar, N., Mehta, G, "Student's academic performance prediction in academic using data mining techniques", In Proceedings of the International Conference on Innovative Computing & Communications (ICICC), **2020.**

[15] Rai, S., Shastry, K. A., Pratap, S., Kishore, S., Mishra, P., Sanjay, H. A, "Machine learning approach for student academic performance prediction", In the proceedings of the Evolution in Computational Intelligence: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Vol.**1**, pp.**611-618, 2020.**