Article title: Improving Students’ Retention Using Machine Learning: Impacts and Implications
Authors: Sandeep Trivedi[1]
Affiliations: Member of IEEE.org, Technocrats Institute of Technology, MP, India[1]
Orcid ids: 0000-0002-1709-247X[1]
Contact e-mail: sandeep.trived.ieee@gmail.com

Preprint statement: This article is a preprint and has not been peer-reviewed, under consideration and submitted to ScienceOpen Preprints for open peer review.
DOI: 10.14293/S2199-1006.1.SOR-.PPZMB0B.v2
Preprint first posted online: 13 August 2022
Keywords: Data Mining, Retention, Machine Learning, Neural Networks, SVM

IMPROVING STUDENTS' RETENTION USING MACHINE LEARNING: IMPACTS AND IMPLICATIONS

Sandeep Trivedi,

Member of IEEE, Technocrats Institute of Technology, Contact: sandeep.wipro2011@gmail.com

Abstract—

Many enrollment management systems rely heavily on student retention. It has an impact on university rankings, school reputation, and financial stability. Student retention has risen to the top of the priority list for higher education administrators. Improving student retention begins with a full grasp of the factors that contribute to attrition. This understanding is the foundation for effectively anticipating at-risk pupils and responding appropriately to retain them. For some years, machine learning approaches have been used in education to predict retention and discover factors impacting retention rates, with better outcomes since 2010. This study focuses on different machine learning techniques used in literature for improving students' retention; we have identified various factors that might affect the students' retention and employed SVM and Neural Networks for predicting students' retention rates. The review presents a research viewpoint on predicting student retention using machine learning through numerous significant results such as the identification of characteristics employed in previous studies and prediction approaches. These findings may be utilized to create more extensive research to enhance prediction capabilities and, as a result, methods to improve student retention.

Keywords: **Data Mining, retention, Machine Learning, Neural Networks, SVM**

INTRODUCTION

THE COVID-19 pandemic is often known as the coronavirus recession which has cast a severe impression on education as well and without improvements in education to address the requirements of the workforce, the skill gap in our workforce will only expand. Today's professions require greater technical skills and higher degrees than in previous generations, [1]. Today's industries need greater technical skills and higher degrees than in previous generations. Jobs that fill a void often demand a college diploma, while individuals without additional skills have struggled, [2].

These forces drive the higher education system to look inward for answers to problems such as rising costs, inequalities, retention, and completion rates. According to statistics by The National Student Clearinghouse Research Center, 65.7% of students at four-year public universities graduate in six years, whereas a dramatic decrease with 39.2% of students at two-year public institutions graduates in three years, [3]. Dropouts are nearly twice as likely to be unemployed as college graduates, and they are four times more likely to default on student loans, damaging their credit and limiting their employment possibilities.

To fulfill the potential of an educated and technically skilled workforce, higher education must assess its methods. The growing number of students who attend community colleges, or two-year schools, during their first two years has necessitated the inclusion of this significant factor in the study. Machine learning techniques have been used to evaluate student data in recent years, which corresponds intention to enhance data processing through data mining, using methods such as neural networks (NN) and support vector machines,[4]. Delen, [5] shows, through several comparison studies, NN, SVM, and decision trees (DT) have better prediction results than other statistical techniques such as logistic regression (LR) and discriminant analysis (DA). Because of their capacity to predict a result using both quantitative and qualitative/categorical data, these approaches have attracted a lot of attention in the literature.

Supervised learning is a machine learning approach that uses training data to construct a computer model through repeated changes to eliminate error. SVM is a supervised learning model that can be used to predict and classify data. A training algorithm is created, which categorizes updated data. SVM is especially effective for determining data clustering in groups. Another predictive modeling method is DT, which uses classification trees to form judgments about a target value based on observations. When the objective is a classification, DTs are employed; however, when the target is a continuous variable, a regression tree is used. When the target is a binary dependent variable, LR is also employed to create a statistical model. Finally, as DA examines data to predict a categorical, dependent variable, it is widely employed in retention research.

Alkhasawneh and Hargraves, [6] created a paradigm that included both qualitative and quantitative research. The factors affecting retention rates were discovered in each research. The essential parameters were then put into a NN model to predict first-year retention rates for students pursuing degrees in science, technology, engineering, and mathematics (STEM). The first research was a quantitative model designed to identify characteristics that have the greatest influence on student retention. The dataset included 1,996 registered students who were divided into two groups: 1,468 registered students and 498 registered minority students. To improve learning time and reduce repetition while feeding the cohorts, the genetic algorithm was utilized to pick factors that had a greater influence on retention. The second research was qualitative, with data gathered through an eight-question survey from a focus group. Content analysis was employed in this section since it is a methodology that is commonly used for textual content. The findings of the two trials were combined into a NN that was run independently to predict GPA and identify whether or not students would be retained.

When employing datasets including all students, the majority of students, and under-represented students, the NN demonstrated overall classification accuracy of 74 percent, 79 percent, and 60 percent, respectively. Furthermore, in the quantitative model, reducing the number of variables for each database enhanced classification accuracy. The research concluded the following fac-tors were useful for predicting performance and retention: first math course grade, high school rank, the impact of pre-college intervention programs, and SAT math score.

This research gives a thorough review of machine learning strategies for improving educational institution retention rates. It provides a research perspective related to the identification of student retention using Machine Learning through previous studies and approaches used for prediction. The review will attempt to address the following research questions: (1) what are the Machine Learning techniques being utilized to forecast student retention rates, and (2) which techniques have demonstrated superior performance in certain settings? , (3) what variables impact the forecast of higher education completion rates?, and (4) what are the issues with predicting student retention? We, mainly, have investigated the use of SVM and Neural networks for predicting students' performance and improving retention.

Different Machine Learning classifiers are studied for evaluating students' retention, [6]. Table.1 shows the methodologies that have been employed for predicting retention over years.

Method	**Study**	**Performance**
Decision trees	Raju & Schumacker (20 I 5)	73.50% 73.75%
Bayesian belief network	Slim et al. (2014)	MSE curves
	Dissanayake et al. (2016) Miranda & Guzman (20 I 7) Uddin & Lee	76% Accuracy was not MSE reported

	(20 I7)	MSE curves
Support vector machines	Slim et al. (2014) McAleer & Szakas (20 I 0) Oztekin (20 I 6) Babic (20 I7)	79.5% 77.6% 57.6%
SVM+DT+NN (ensamble)	Adejo & Connolly	81.6%
K-nearest neighbor	Dissanayake et al. (2016)	83.3%
	lam-On & Boongoen (2017)	93.3%
Logistic regression	Delen (20 I1) Kondo et al. (2017)	74.3% 75%
Neural Networks	Delen (20 I1) Raju & Schumacker (20 I 5) Miranda & Guzman (20 I 7) Adejo & Connolly	79.8% 77.7% 83% 73%
Linear Support vector machines (SVM)	Naicker, N., Adeliyi, T., & Wing, J. (2020)	N/A
Deep Learning Approach (NN)	Agrusti, F., Mezzini, M., & Bonavolontà, G. (2020).	N/A

Table.1 Methodologies employed for prediction over years

Various studies have been proposed in the literature for improving students' retention. A comparative study is provided in the following table.2:

Ref. and publication year	Focus/Scope of Survey	Type of Review	Years covered	Strengths	Limitations
(2022) [30]	Educational Data mining and Predictive Analytics in three learning environments	Systematic review	(2017-2021)	Identifies the success factors and the features that are not indicative of predicting a student's performance	Reviews article since 2017
2021 [31]	ML techniques to predict performance	Systematic review	(2009-2021)	Well formulated questions	No comparison of the reviewer articles
2021 [32]	Data mining and LA techniques to predict student's performance	Systematic review	(2010-2020)	A comprehensive review with practical limitations	No information regarding the factors used in the prediction
2020 [33]	Prediction of academics in higher education	Traditional review	(2016-2020)	Provides a holistic view and review of applying predictive techniques	Only presents the influential factors
2020 [34]	ML techniqu	Traditi	(2015-	Provides a	Does not

	es to analyze and predict student performance	onal review	2020)	summarized table of the main features of the reviewed articles	include details about the used factors
2019 [35]	Predictive modeling technique for monitoring student's performance	Traditional review	NA	Reviewed different articles for predicting performance	No comparative analysis of the articles
2018 [36]	Predicting academic performance	Systematic review	(2010-2017)	Explains the drawbacks from the perspective of predicting performance	Does not identify the feature that is not indicative of predicting performance
2017 [37]	Using data mining for student performance prediction	Literature review	(2007-2016)	Presents predictive data mining techniques focus on successful factors	A limited number of papers reviewer don't include literature's limitation
2017 [38]	Predicting student's performance using LA	Systematic review	(2002-2016)	Classifies predictive models and articles according to methodologies	The preliminary study lacks LA techniques and other salient details
Current review	Predicting student retention using ML	Systematic + traditional review	(2002-2022)	Provides a detailed literature overview along with methodologies and results	

Table.2. Comparative analysis of the current state-of-the-art techniques

2. PREDICTING STUDENT RETENTION USING SUPPORT VECTOR MACHINES (SVM)

SVM is categorized as a supervised learning algorithm that performs regression or classification for numeric responses and categorical variables. It creates a mapping space to separate the input data into different classes. SVM maps both linear and non-linear data by using kernel functions to transform the inputs to a higher-dimensional space, which allows for a linear separability The usage of kernels, therefore, lowers the cost.

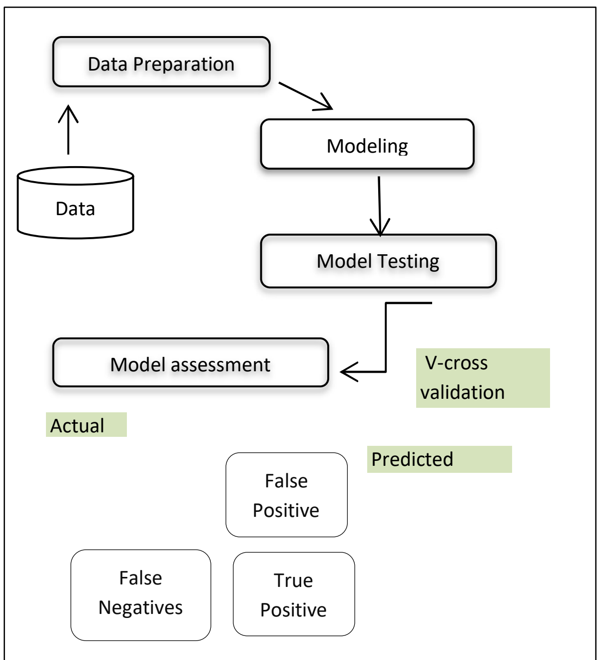
By separating the data into parallel hyperplanes, you can reduce the problem's complexity. The optimum condition is found by minimizing the Euclidean norm of the weight vector, which is a constrained optimization problem that can be solved using the method of LaGrange multipliers.

The program aims to optimize the margin between parallel hyperplanes, which limits misclassification. As the distance between the hyperplanes grows, it is predicted that the generalization error decreases.

Research methodology using SVM

The research process/methodology is conducted in the following phases shown in figure.1. i.e., 1) data description and preparation, 2) data modeling, application of SVM, and 3) model assessment.

The dataset was made up of 904 students who were pursuing degrees in chemistry, biology, or engineering over five years. Based on this information, 177 students were found to have completed their degree in less than three years, which is 150 percent of the usual time for completion as defined by the 1990 Student Right-to-Know Act for postsecondary institutions. The other 727 students did not complete their degree within that time frame, owing to causes such as college dropout or changing their major to something other than STEM. Due to a large amount of missing data and inconsistent data, the data set was cleansed. Some students' standardized test scores, for example, were absent since this information is not necessary for community college entrance. After cleansing the data to remove any du



Data description and preparation

Figure.1. Methodology using SVM

INITIAL INPUT VARIABLES	
X_1	Degree
X_2	Gender
X_3	Age
X_4	1st student
X_5	1st generation
X_6	ACT English
X_7	ACT composite
X_8	ACT Math
X_9	ACT Reading
X_{10}	High school GPA
X_{11}	Plans to work
X_{12}	College GPA
X_{13}	FT student

Table.3. Initial input variables

INPUT VARIABLES (after filtering)	
X_1	Degree
X_2	Gender
X_3	Age
X_4	1st generation
X_5	ACT composite
X_6	High school GPA
X_7	Plans to work
X_8	College GPA

Table.4. Variables used in the model

Model

SVM type 2 classification was used as the model. For a discrete target variable, this approach classifies binary data. The classifier uses the radial basis function (RBF), which is also known as the kernel for dimensional transformation. The prediction model was trained and validated using k-fold cross-validation. A 0.01 error objective and a maximum number of iterations of 10,000 were defined as a halting criterion.

Model Specification	Value
SVM type	Classification type-2
Kernel type	Radial basis function
No. of SVs (0)	34
No. of SVs (1)	48
No. of independent variables	8

Table. 5. Model summary

Model assessment

In the validation set, precision and recall metrics were used to test the model, as well as overall accuracy. By avoiding possible misinterpretations, the last phase provides a more comprehensive study of the data. The model must be accurate in predicting non-completers (low error type II) because the data will be used to enhance and build retention tactics, which will cost the institution money if they invest in students who are a false negative for completion risk. The overall performance was calculated as the proportion of correctly classified values from the training, testing, and validation subsamples obtained from the k-fold cross validation application.

Results

Because the model employed soft bounds, bounded vectors are placed within the margin area. Only 9% of the categorized vectors are bounded, which indicates a solid model implementation since data generalization is better when the number of bounded vectors is low concerning the total cases. Table.6. Summarizes the results and shows that the model can classify with an accuracy of above 70% with modest misclassification (false positive). In addition, the model is more accurate when it comes to predicting non-completers.

Class	0	1	Total	Recall
0	39	8	47	0.82
1	7	17	24	0.70
Total	46	25	71	
Precision	0.84	0.68		

Table.6. Confusion matrix, recall, and precision measures

Despite the lack of weights used to prioritize class categorization, the results are more accurate in identifying students who are in danger of dropping out. This is vital to keep in mind while developing retention tactics that depend on purposeful counseling, as correcting false-positive misclassifications can be costly. This is why the recall measure is the focus of the model study.

The model's overall accuracy is high, as seen in Table 7. However, there is a clear distinction between training and testing results.

Classification Accuracy (%)	
Train	94.3
Test	78.8
Overall	90.4

Table.7. Accuracy

The testing accuracy provides more information about the prediction performance in this scenario since it eliminates misinterpretations due to data overfitting. The model then has a strong prediction performance when the testing accuracy is greater than 78 percent, which is a sufficient metric for the problem's prediction aims.

3. Neural Networks

Neural networks (NN) have been widely employed in technical applications involving prediction and categorization in recent decades, particularly in engineering, business, and medicine. Because of its importance and effectiveness, the neural network model is particularly appealing for modeling complex systems: universal function approximation capacity, accommodation of several non-linear variables with unknown interactions, and strong generalization ability, [6].

Few studies have been published on the application and accuracy of data-mining tools in institutional research, [8], [9]. The use of neural networks and decision tree analysis in predicting community college students' transfer to four-year colleges was shown, with the conclusion that a classification and regression tree (C&RT) method achieved overall better accuracy than decision trees, [10]. Byers González and DesJardins [7] found that neural networks outperformed binary logistic regression in predicting the application behavior of potential freshmen who submitted admission test results to a prominent research university.

Although cumulative research on time to degree (TTD) completion is less spectacular, regression and route analysis models have contributed significantly to our knowledge of student retention, [11, 12, 13]. The more complicated nature of the road to graduation, which has stretched significantly over the previous thirty years for a bachelor's degree, is a plausible cause.

Methodology

The neural network model is particularly interesting for modeling complex systems because of its importance and effectiveness: universal function approximation capability, accommodation of numerous non-linear variables with unknown interactions, and excellent generalization ability. The *overall college GPA* is the response variable in this research.

In this study, two different models were created utilizing a multilayer feed-forward backpropagation network as illustrated in Figure.2, where we have three layers i.e., input layer, hidden layer, and output layer to 1) predict incoming freshmen retention and 2) predict incoming freshmen retention. 2) Divide the same group into three groups: at-risk, intermediate, and advanced. At-risk students have a lower GPA and are more likely to drop out of an S&E major, whereas advanced students have a better GPA and are less likely to drop out. Students were split into three categories depending on their total GPA: at-risk, intermediate, and advanced. At-risk students had a GPA of less than 2.7; intermediate students had a GPA of 2.7–3.4, and advanced students had a GPA of more than 3.4. The average GPA categorization in higher education was used to create this classification.

The network contains a six-element input layer, an eleven-element hidden layer, and a single-element output layer.

The network that solves non-linear least squares problems was trained using the Gauss-Newton learning approach. S&E majors were modeled separately using 8 and 6 elements, respectively.

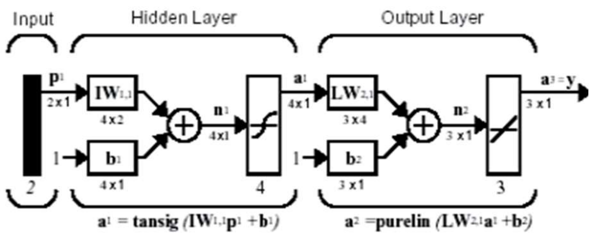


Figure.2. Multilayer Feed-Forward Backpropagation Network [14]

RESULTS

To minimize overfitting, the 10-fold cross-validation technique was proposed for all S&E majors . The absolute GPA prediction model has an r-value of 0.54 and an accuracy range of [-0.5, 0.5]. The margin of error for [0.5] is 68%.

The actual predicted GPA plot is shown in figure.3 which shows that the absolute GPA for science majors yielded better results compared to both engineering and science.

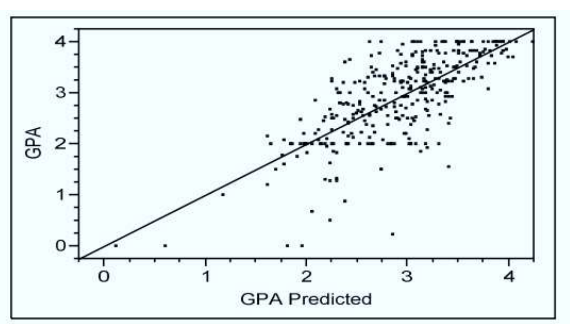


Figure.3 Regression analysis of actual GPA by Predicted GPA plot

Table.8. which is a further elaboration of figure.3., shows that science majors alone showed better results than predicting both science and engineering or engineering alone. The mean for forecasting S&E is roughly 0.42, 0.40 for predicting science, and 0.41 for engineering, according to an analysis of the model's error within the [-0.5, 0.5] margin of error. This is a good indication of the model's accuracy.

Variable	S&E	Science	Engineering
R-value	0.54	0.57	0.59
Accuracy	68%	70.5%	68.9%
Total	338	190	148

Table.8. Accuracy by Major and outcomes by r

Variable	S&E	Science	Engineering
Min	0.002808	0.000519	8.06E-05
Max	2.6238	1.6528	2.7725
Mean	0.427	0.407178	0.41065

Table.9. Summary of the result of errors

In terms of the classification model, 70.1% of the output was properly categorized, with an R-value of 0.41. In figure.4., the Receiver Operating Characteristics (ROC) graph shows that the concept of dividing incoming freshmen into three categories: at-risk, moderate, and advanced appears to be a viable test. For advanced, intermediate, and at-risk pupils, the area under the curve is 91.5 percent, 87.2 percent, and 85.8%, respectively. In addition, 10-fold cross-validation was employed.

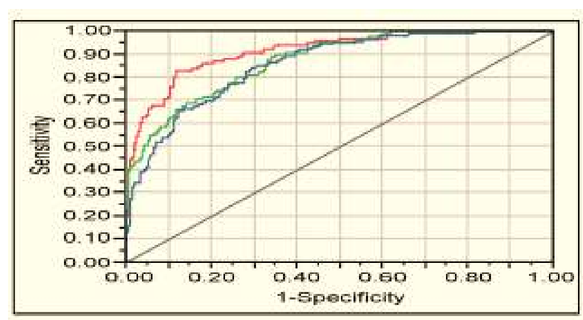


Figure.4 Regression analysis of actual GPA by Predicted GPA plot

When it comes to forecasting incoming freshman retention at VCU S&E disciplines, the findings of the two models shown above are highly encouraging. The outcomes of neural networks might be greatly improved with big data sets (i.e. more than 500 students), according to the literature [4]. For our sample size and constrained parameters, a prediction accuracy of 68 percent for absolute GPA is reasonable. Furthermore, the ROC curve provided a solid indicator of how well our classification model worked, and it is thought that results might be enhanced with a bigger data set and additional associated criteria like arithmetic performance.

In this data collection, 70 pupils were identified as African American, Hispanic American, or Native American minorities. Due to the small number of minority students in this data set and the restricted number of variables available, race was used as an input variable rather than comparing the performance of minority and majority groups. The research was also confined to the 2008 academic year. Future research might incorporate a more diversified data collection with a higher proportion of minority students.

When predicting the success of science students, the results were marginally better than when predicting the performance of engineering students. There is no adequate rationale for this disparity at this moment, although it is speculated that it is connected to the variety in input factors for scientific students.

CONCLUSION

For decades, many scholars have been concerned about attracting more students to scientific and engineering areas. This paper gives a thorough assessment of the research on using machine learning algorithms to predict student retention in higher education using variables such as dropout risk, attrition risk, and completion risk. We have provided a comparative analysis of the current studies used in predicting student retention. The proposed article not only reviews the current state-of-the-art techniques of ML for predicting student retention but also addresses in detail the important factors used for prediction. The proposed article explains the detailed methodology used in SVM and Neural Networks along with the results and findings. The main contribution of the proposed article is *1)* A holistic review of current techniques to predict student retention; *2)* Challenges in predicting student retention; *3)* Exploring SVM and Neural Networks to predict retention with results (implications and applications).

We have investigated SVM and Neural networks in predicting students' retention. At epoch 2919, the SVM's best performance was attained with an error of 0.01. In other words, the model achieved its error objective and terminated training. SVM can classify with an accuracy of above 70% with modest misclassification (false positive). In addition, the model is more accurate when it comes to predicting non-completers.

Neural network approaches will be used to model S&E incoming freshmen retention, paving the door for a greater understanding of student retention characteristics. The models established here are intended to forecast absolute GPA and to categorize students into three groups depending on their total GPA: at-risk, intermediate, and advanced. We have observed that Neural Networks' accuracy is more than SVM and hence it can classify the performance of students, or students' retention more precisely.

The study had a small sample size and a small population, and it was intended to predict college retention in general. Future research will be tailored to the student's grade level (freshman, sophomore, etc.) and race/ethnicity.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

The author received no financial support for the research, authorship, and/or publication of this article.

REFERENCES

1. Rosendale, J., & Wilkie, L. (2020). Scaling workforce development: using MOOCs to reduce costs and narrow the skills gap. *Development and Learning in Organizations: An International Journal*.

2. Carnevale, A. P., Jayasundera, T., & Gulish, A. (2016). America's Divided Recovery: College Haves and Have-Nots. *Georgetown University Center on Education and the Workforce*.

3. Shapiro, D., Dundar, A., Huie, F., Wakhungu, P. K., Bhimdiwala, A., & Wilson, S. E. (2018). Completing college: A national view of student completion rates – fall 2012 cohort (Signature Report 16). National Student Clearinghouse Research Center

4. Cardona, T., & Cudney, E. (2019). Predicting student retention using support vector machines [Paper presentation]. Proceedings of the International Conference on Production

Research Manufacturing Innovation: Cyber-Physical Manufacturing, Chicago, IL, United States... Proceedings of the International on Production Research Manufacturing Innovation: Cyber-Physical Manufacturing, Chicago, IL, United States."]

5. Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498-506.

6. Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a hybrid model to predict student first year retention in STEM disciplines using machine learning techniques. *Journal of STEM Education: Innovations and Research*, *15*(3).

7. González, J. M. B., & DesJardins, S. L. (2002). Artificial neural networks: A new approach to predicting application behavior. *Research in Higher Education*, *43*(2), 235-258.

8. Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498-506.

9. Herzog, S. (2006). Estimating student retention and degree-completion time: Decision Trees and Neural Networks vis-à-vis Regression. *New directions for institutional research*, *131*, 17-33.

10. Mason, C., Twomey, J., Wright, D., & Whitman, L. (2018). Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression. *Research in Higher Education*, *59*(3), 382-400.

11. Adam, A. J., & Gaither, G. H. (2005). Retention in Higher Education: A Selective Resource Guide. *New Directions for Institutional Research*, *125*, 107-122.

12. Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research. Volume 2*. Jossey-Bass, An Imprint of Wiley. 10475 Crosspoint Blvd, Indianapolis, IN 46256.

13. Braxton, J. M., Milem, J. F., & Sullivan, A. S. (2000). The influence of active learning on the college student departure process: Toward a revision of Tinto's theory. *The journal of higher education*, *71*(5), 569-590.

14. https://www.mathworks.com/help/deeplearning/backpro3.html#48640

15. Ahmad, Z., & Shahzadi, E. (2018). Prediction of Students' Academic Performance Using Artificial Neural Network. *Bulletin of Education and Research*, *40*(3), 157-164.

16. Raju, D., & Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models. *Journal of College Student Retention: Research, Theory & Practice*, *16*(4), 563-591.

17. Slim, A., Heileman, G. L., Kozlick, J., & Abdallah, C. T. (2014, December). Predicting student success based on prior performance. In the *2014 IEEE symposium on computational intelligence and data mining (CIDM)* (pp. 410-415). IEEE.

18. Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016). Predictive modeling for student retention at St. Cloud state university. In *Proceedings of the International Conference on Data Science (ICDATA)* (p. 215). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

19. Mendoza, K. L. G., Heras, J. R., Mendoza, Y. F., & Torres, Z. C. F. (2022). Deserción estudiantil en Colombia: un análisis de los factores predictores y sus indicadores. *Encuentros. Revista de Ciencias Humanas, Teoría Social y Pensamiento Crítico.*, (Extra), 297-310.

20. Uddin, M. F., & Lee, J. (2017). Proposing stochastic probability-based math model and algorithms utilizing social networking and academic data for good fit students prediction. *Social Network Analysis and Mining*, *7*(1), 1-21.

21. Sepulveda, T. A. C. (2020). *Development of a system architecture for the prediction of student success using machine learning techniques*. Missouri University of Science and Technology.

22. Babić, I. Đ. (2017). Machine learning methods in predicting the student academic motivation. *Croatian Operational Research Review*, 443-461.

23. Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*.

24. Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016). Predictive modeling for student retention at St. Cloud state university. In *Proceedings of the International Conference on Data Science (ICDATA)* (p. 215). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

25. Lam-On, N., & Boongoen, T. (2014, December). Using cluster ensemble to improve classification of student dropout in Thai university. In *2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (scis) and 15th International Symposium on Advanced Intelligent Systems (isis)* (pp. 452-457). IEEE.

26. Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, *13*(1), 17-35.

27. Takahara, M., Shundo, M., Kondo, M., Suzuki, K., Nambu, T., & Ogino, T. (1998). Early detection of osteochondritis dissecans of the capitellum in young baseball players. Report of three cases. *JBJS*, *80*(6), 892-7.

28. Naicker, Nalindren, Timothy Adeliyi, and Jeanette Wing. "Linear support vector machines for prediction of student performance in school-based education." *Mathematical Problems in Engineering* 2020 (2020).

29. Agrusti, Francesco, Mauro Mezzini, and Gianmarco Bonavolontà. "Deep learning approach for predicting university dropout: A case study at Roma Tre University." *Journal of e-Learning and Knowledge Society* 16.1 (2020): 44-54.

30. Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student Retention using Educational Data Mining and Predictive Analytics: A Systematic Literature Review. *IEEE Access*.

31. Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552.

32. Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, *11*(1), 237.

33. 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.

34. Liz-Domínguez, M., Caeiro-Rodríguez, M., Llamas-Nistal, M., & Mikic-Fonte, F. A. (2019). Systematic literature review of predictive

analysis tools in higher education. *Applied Sciences*, *9*(24), 5569.

35. Asiah, M., Zulkarnaen, K. N., Safaai, D., Hafzan, M. Y. N. N., Saberi, M. M., & Syuhaida, S. S. (2019). A review on predictive modeling technique for student academic performance monitoring. In *MATEC Web of Conferences* (Vol. 255, p. 03004). EDP Sciences.

36. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., ... & Liao, S. N. (2018, July). Predicting academic performance: a systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education* (pp. 175-199).

37. Kumar, M., Singh, A. J., & Handa, D. (2017). Literature survey on student's performance prediction in education using data mining techniques. *International Journal of Education and Management Engineering*, *7*(6), 40-49.

38. Hu, X., Cheong, C. W., Ding, W., & Woo, M. (2017, March). A systematic review of studies on predicting student learning outcomes using learning analytics. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 528-529).