Predicting Student Academic Performance Using Machine Learning

Chapter *in* Lecture Notes in Computer Science · September 2021

DOI: 10.1007/978-3-030-87013-3_36

8 authors, including:

Opeyemi Ojajuni
Southern University and Agricultural and Mechanical College
7 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Foluso Ayeni
University of Nebraska at Omaha
34 PUBLICATIONS **199** CITATIONS
SEE PROFILE

Femi Ekanoye
ICT University
8 PUBLICATIONS **65** CITATIONS
SEE PROFILE

Sanjay Misra
Institute for Energy Technology
852 PUBLICATIONS **11,199** CITATIONS
SEE PROFILE

Predicting Student Academic Performance Using Machine Learning

Opeyemi Ojajuni[1], Foluso Ayeni[2](✉), Olagunju Akodu[3], Femi Ekanoye[4], Samson Adewole[4], Timothy Ayo[4], Sanjay Misra[5], and Victor Mbarika[6]

[1] Department of Science and Mathematics Education, Southern University and A&M College, Baton Rouge, USA
Opeyemi_ojajuni_00@subr.edu

[2] Department of Information Systems and Quantitative Analysis, University of Nebraska, Omaha, USA
fayeni@unomaha.edu

[3] Department of Electrical and Electronics Engineering, Southern University and A&M College, Baton Rouge, USA
olagunju_akodu_00@subr.edu

[4] Global Technology Management and Policy Research Group, Southern University and A&M College, Baton Rouge, USA
femi_ekanoye@subr.edu, oluwadamilaresam@gmail.com, timothyayo99@gmail.com

[5] Department of Information and Communication Engineering, Covenant University, Ota, Nigeria
sanjay.misra@covenantuniversity.edu.ng

[6] Department of Management Information Systems, East Carolina University, Greenville, USA
mbarikav20@ecu.edu

Abstract. The introduction of the Internet of Things (IoT), Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and Big Data have paved the way for research focused on improving the student learning experience and help to address challenges faced by the education system. Machine Learning technology analyzes data to recognize patterns and use them to make predictions. This paper introduces a ML model that classify and predict student academic success by utilizing supervised ML algorithms like Random Forest, Support Vector Machines, Gradient boosting, Decision Tree, Logistic Regression, Regression, Extreme Gradient Boosting (XGBoost), and Deep Learning. This paper aims to predict student's academic success based on historical data and identify the key factors that affect student academic success. Thus, the proposed approach offers a solution to predict student academic performance efficiently and accurately by comparing several ML models to the Deep Learning model. Results show that the Extreme Gradient Boosting (XGBoost) can predict student academic performance with an accuracy of 97.12%. Furthermore, results showed significant social and demographic features that affect student academic success. This study concludes that applying Machine Learning technology in the classroom will help educators identify gaps in student learning and enable early detection of underperforming students, thus empowering educators with informed decision-making.

O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12957, pp. 481–491, 2021.
https://doi.org/10.1007/978-3-030-87013-3_36

Keywords: Machine learning · Deep learning · Student academic performance · Educational data mining · Data analytics · Convolutional Neutral Networks (CNN)

1 Introduction

Educational data mining (EDM) applies data mining, machine learning, and deep learning to data generated in an academic setting to improve student learning experiences [1, 2, 3]. The interaction of students with learning platforms and materials creates large amounts of data [4, 5]. Analyzing this data provides insight into the student learning process and student achievement. Further analysis can identify academic, demographic, and social factors affecting student academic success. Student academic success is measured by assessing student performance across academic subjects. Teachers measure student academic performance from different approaches, ranging from students' final grades, Grade Point Average (GPA), and Standardized Tests. According to reports from the United States of America Department of Education and National Assessment of Educational Progress (NAEP), the education system suffers from several challenges like student academic underachievement, increased university dropout rates, graduation delays, and inadequate student workforce readiness. Over the years, student academic success has continued to decline, even more prevalent amongst minority students [6, 7, 8]. Education technology advancements such as Artificial Intelligence (AI), Virtual Reality (VR), 3D printing, smart multimedia devices, Internet of Things (IoT), and Machine Learning are beginning to improve the student learning process and management [9].

Machine Learning analyzes data to recognize patterns and use those patterns to make predictions. Applying ML in the classroom will enable educators to identify critical factors affecting student's success. Furthermore, ML will allow educators to identify underperforming students, thus empowering educators with informed decision-making. Several tools such as R Software, Python Scikit-learn, TensorFlow are currently used in ML technology. A wide range of ML algorithms is also available for predicting student academic performance. These algorithms include Random Forest, Support Vector Machines (SVM), AdaBoost, Decision Tree, Naive Bayes, and K-nearest Neighbors.

In this research work, we aim to use historical education data on student academic performance collected from the UC Irvine Machine Learning Repository to identify the key factors that affect student academic achievement. Furthermore, the research intends to predict future student academic success by recognizing patterns in the historical dataset and using the patterns to make predictions. The research objectives addressed in this research work are listed below:

1. What are the factors that have significant effect on students' academic success?
2. How can these factors predict student academic performance using machine learning?

The research paper is organized under the following subheading: Related research work, methods and implementation, results, and conclusion.

2 Related Research Work

Learning management systems have empowered education institutions with interactive learning tools such as game-based, simulation applications, virtual reality, and e-learning systems. These platforms have allowed researchers to collect and analyze student data [2, 5]. The authors [9] applied the Decision Tree, Neural Network, and Support Vector Machine (SVM) classification ML algorithm to predict academic performance from student internet usage behaviors. Their results showed that student internet usage behaviors effectively predict academic performance with an accuracy of 71%–76%; however, the authors only considered accuracy as the performance metric. In [10] work, the authors proposed a system that uses ML algorithms trained to predict students' academic performance by classifying them into bad or good. The model was trained on data gathered from a university source and implemented using the K-nearest neighbor and Decision tree classifier. The result showed that the Decision tree classifier has 94.44% accuracy, but the author considered only accuracy as its performance metrics.

Similarly, the authors [2] proposed a classification ML model using SVM and Logistic regression classifiers to predict students' academic performance. The model extracted features from the preprocessed dataset obtained from an online educational platform to classify student academic performance as bad, average, or good. The result showed that the SVM produced an accuracy of 79%, which was higher than the logistic regression. The authors considered accuracy, recall, precision, and f1-score using confusion box metrics to evaluate the system's performance. The authors [1] used Naïve Bayes, Random Forest classifier, and Ensemble learners classification ML model to predict student academic performance using a dataset comprising 887 instances of 19 attributes of first-year students. The Random Forest classifier outperformed other models with an accuracy of 93%. Evaluation metrics of recall, precision, and f1-score using confusion box metrics was employed in evaluating the model performance. Research on ML in education is still in its preliminary stages, there are still many challenges such as prediction accuracy, overfitting, underfitting, deployment of the model that need attention. Thus, our proposed approach offers an efficient and accurate student academic performance by comparing several ML models to deep learning models. Generally, deep learning models have better accuracy because they extract features from the dataset in an incremental manner. ML algorithms are applied to the dataset to analyze and identify features that significantly impacted student academic performance. Finally, leveraging these features, several ML models are trained to classify and predict student academic performance category, and we also compared the model's performance based on accuracy score and cross-validation score.

3 Material and Methods

3.1 Tools

The experiments were conducted on a computer running MacOS Big Sur operating system with the specification of 2.3 GHz Dual-Core Intel Core i5 with 8 Gigabytes memory. Python programming language was used along with Scikit-learn, and TensorFlow ML libraries to implement algorithms, build ML model, and obtain statistical results [11, 12].

3.2 Dataset

The dataset used in this study was from the UC Irvine Machine Learning repository [13]. The dataset consists of 1044 student's academic performance in two high schools. The data attributes include demographic, social, and academic related features. Table 1 shows the summary of our dataset attributes.

Table 1. Dataset [13] attributes

Feature category	Name of the attributes	Description	Attribute type
Demographical features	School	Student's school	Categorical
	Sex	Student's sex	Categorical
	Age	Student's age	Numeric
	Address	Student's home address type	Categorical
	Famsize	Family size	Categorical
	Pstatus	Parent's cohabitation status	Categorical
	Medu	Mother's education	Numeric
	Fedu	Fedu - father's education	Numeric
	Mjob	Mother's job	Categorical
	Fjob	Father's job	Categorical
	Reason	Reason to choose this school	Categorical
	Guardian	Guardian - student's guardian	Categorical
Social features	Internet	Internet access at home	Categorical
	Romantic	With a romantic relationship	Categorical
	Famrel	Quality of family relationships	Numeric
	Freetime	Free time after school	Numeric
	Goout	Going out with friends	Numeric
	Dalc	Workday alcohol consumption	Numeric
	Walc	Weekend alcohol consumption	Numeric

(*continued*)

Table 1. (*continued*)

Feature category	Name of the attributes	Description	Attribute type
	Health	Current health status	Numeric
Academic related features	Absences	Number of school absences	Numeric
	Traveltime	Home to school travel time	Numeric
	Studytime	Weekly study time	Numeric
	Failures	Number of past class failures	Numeric
	Schoolsup	Extra educational support	Categorical
	Famsup	Family educational support	Categorical
	Paid	Number of past class failures	Numeric
	Activities	Extra-curricular activities	Categorical
	Nursery	Attended nursery school	Categorical
	Higher	Wants to take higher education	Categorical
	Final grade	Final grade	Numeric

3.3 Data Preprocessing and Feature Engineering

Data preprocessing is done on the dataset to check for null values, duplicates, and invalid values. Fortunately, our dataset is clean and ready for encoding. The final grade was converted into multiclass categories- “excellent, good, satisfactory, poor, and failure” under the following conditions:

- Excellent – final grade score is between 45–60
- Good– final grade score is between 36–44
- Satisfactory– final grade score is between 24–35
- Poor – final grade score is between 20–23
- Failure – final grade score is between 0–23

ML models require all input and output data to be attributed to numeric values. Any data that is not numeric must be encoded to numeric values before fitting it into a ML model. Several attributes are non-numeric and categorical in our dataset, as seen in Table 1. This study employs the One-Hot-encoding in Python's Scikit-Learn to encode and normalize non-numeric and categorical data attribute type [11]. Feature engineering techniques help in extracting important features from the dataset.

3.4 Machine Learning Classification Model

Solving problems with ML is grouped into supervised and unsupervised learning. Unsupervised ML works with unstructured data, while supervised ML works with a structured dataset where the input variables are mapped with the output variables. Supervised ML problems are grouped into regression and classification problems [14]. Regression problems involve predicting a continuous, discrete value, for example, predicting student final grade score. ML classification refers to the process of predicting a category from input data points. The category output can be binary classification - "fail" or "pass" or multi-class classification- "excellent, good, satisfactory, poor, and failure". ML classification is a supervised ML where input data is labeled and mapped with the output data; the ML model lis trained to predict the output from input. Implementing a ML classifier requires importing the necessary ML module package, then loading the dataset [14]. Data pre-processing and cleaning are done on the dataset to check for null values, duplicates, invalid values and encode non-numeric and category data attribute types.

After successful data preprocessing, the feature engineering technique explores the dataset to understand the correlation relationship between variables to identify features that significantly impact the output variable. This enabled us to improve the model's accuracy by removing attributes that significantly impact the output variable (final student grade) but not an essential feature in predicting student academic performance. The refined dataset is then split into training & testing sets. The training dataset trains the model, and the testing dataset measures the model's performance based on accuracy and cross-validation. Figure 1 shows this study ML model flowchart. This study built and trained the following ML classification algorithms: Random Forest, Support Vector Machine classifier, Stochastic Gradient Descent, Decision Tree, Adaptive Boosting, Logistic Regression, and Deep Learning. Deep learning is a technique that uses neural network concepts to build and train ML models. Deep learning consists of the input layer (receives the input data), hidden layer (incrementally extracts important features), and the output layer [15]. Deep learning consisting of a Convolutional Neural Network (CNN) model with four hidden layers is suitable for our research objectives.

3.5 Machine Learning Model Performance Evaluation

ML uses the testing dataset to measure the performance of the model. Accuracy, cross-validation, precision, recall, F1-score, confusion matrix, log loss, Receiver Operating Characteristic (ROC), and Area Under Curve (AUC) are some of the performance metrics used to evaluate ML classification model [16]. This research employs accuracy and cross-validation as performance metrics to evaluate the ML classification models. The CNN model's performance was evaluated using a confusion matrix to calculate the model's accuracy, precision, and sensitivity. Accuracy is the total number of correct predictions out of the total number of predictions [7]. Cross-validation assesses how effective the model will work on a new dataset. The confusion matrix is an error matrix that virtualizes ML model performance. The confusion matrix is used to calculate the accuracy, precision, and sensitivity of the model. Precision is the ratio of correctly predicted values to total predicted values. Sensitivity evaluates the proportion of correct prediction the model gets right [7].

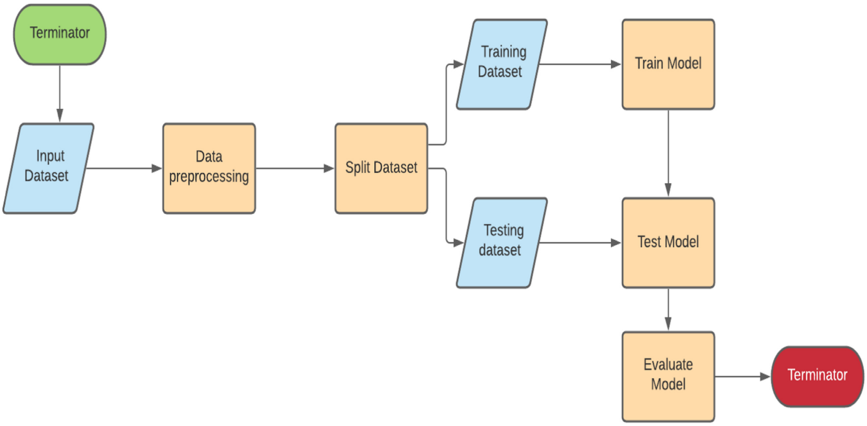


Fig. 1. ML model flowchart

4 Implementation and Result

The "plot_importance" function in Scikit-learn library help in plotting the important features that affect student final grade. In predicting student academic performances, the order of importance of features and its score can be seen in Fig. 2. The number of school absences has the highest importance score. This indicates that students who miss school are more likely to have poor academic performance. Current health status, going out with friends, free time after school, quality of family relationships are major social features that affect student academic performance. Mother's job, father's job, Parent's cohabitation status, student's home address type, and reason to choose this school are the most minor features that affect student academic performance.

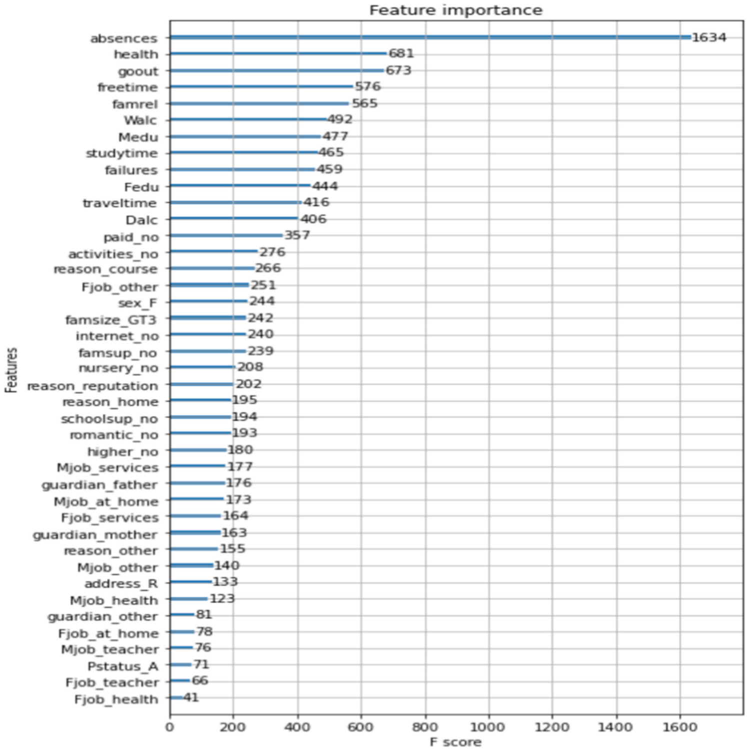


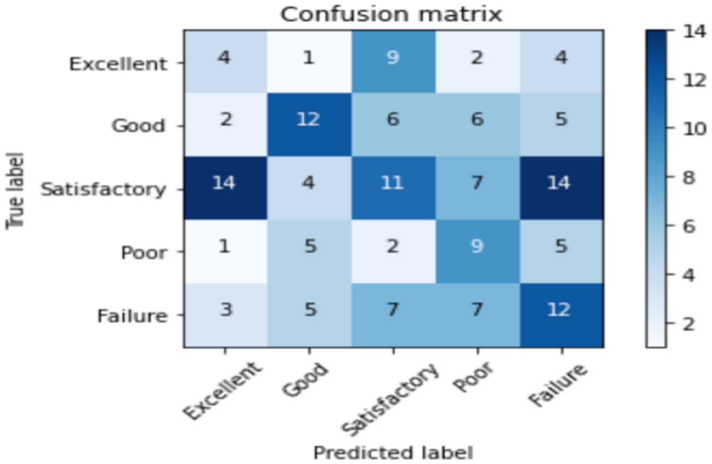
Fig. 2. Important features and its score

To get an accurate evaluation of our model, the dataset containing 1044 students is split into train and test dataset in 70% to 30% ratio using the ‘train_test_split’ function in sci-kit learn. After building and training the ML model, the cross-validation function ‘cross_val_score’ helped compute the model’s average accuracy on the test dataset. The cross-validation function divides the test dataset into smaller subsets. The subsets are then fit into the model and compute the accuracy score five times with different subsets each time [17]. After applying various classification models to the dataset, different accuracy and cross-validation score were obtained for each model. Table 2 shows the accuracy and cross-validation scores for each model. The Deep Learning model gave an accuracy of 72.74%, precision of 30.31%, and sensitivity of 31.38% . Figure 3 shows the confusion matrix used in calculating the performance matrix. The Extreme Gradient Boosting (XGBoost) model outperforms other models in predicting student academic performance. XGBoost Model gave 97.12% accuracy and 35.67% cross-validation. Since the XGBoost model gave the best accuracy, this indicates that the XGBoost ML model is the most suitable ML model considering the nature of our dataset and research objectives.

Table 2. Comparison of Machine Learning models

ML classifier	Accuracy (%)	Cross validation (%)
Decision Tree Model	47.95	30.89
Random Forest Model	92.60	35.66
Support Vector Classifier Model	42.88	34.39
Logistic Regression Model	40.96	36.62
Ada Boost Model	35.75	32.48
Stochastic Gradient Descent	33.69	33.121
XGBoost Model	97.12	35.67

Deep Learning (CNN)	72.22	Precision = 30.31
		Sensitivity = 31.38



	Excellent	Good	Satisfactory	Poor	Failure
Excellent	4	1	9	2	4
Good	2	12	6	6	5
Satisfactory	14	4	11	7	14
Poor	1	5	2	9	5
Failure	3	5	7	7	12

Fig. 3. Deep Learning confusion matrix

5 Conclusion and Future Work

This study has strengthened and explored how Machine learning can empower educators with informed decision-making. Predicting student academic performance or success is an essential concept in tackling the student academic performance crisis. This study used several ML classification models to predict student academic performance. Results showed a range of accuracy from 33% to 98% and a range of cross-validation from 30% to 37%. The XGBoost Model is the most suitable ML model by achieving 97.12% accuracy and 35.67% cross-validation. Furthermore, results showed that the number of school absences, current health status, going out with friends, free time after school, quality of family relationships is significant features that affect student academic performance. This study concludes that this research work can help educators identify gaps in student learning and enable early detection of underachieving students, thus empowering educators with informed decision-making, ultimately improving student academic success and learning process.

References

1. Jayaprakash, S., Krishnan, S., Jaiganesh, V.: Predicting students academic performance using an improved random forest classifier. In: 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, pp. 238–243, March 2020. https://doi.org/10.1109/ESCI48226.2020.9167547
2. Bhutto, E.S., Siddiqui, I.F., Arain, Q.A., Anwar, M.: Predicting students' academic performance through supervised machine learning. In: 2020 International Conference on Information Science and Communication Technology (ICISCT), Karachi, Pakistan, pp. 1–6, February 2020. https://doi.org/10.1109/ICISCT49550.2020.9080033
3. Jacob, J., Jha, K., Kotak, P., Puthran, S.: Educational data mining techniques and their applications. In: 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), pp. 1344–1348, October 2015. https://doi.org/10.1109/ICGCIoT.2015.7380675
4. Al Mayahi, K., Al-Bahri, M.: Machine learning based predicting student academic success. In: 2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Brno, Czech Republic, pp. 264–268, October 2020. https://doi.org/10.1109/ICUMT51630.2020.9222435
5. Olaperi, Y., Fernandez-Sanz, L., Medina, J., Misra, S.: Framework for academic advice through mobile applications (2016)
6. Statement from Secretary DeVos on 2019 NAEP Results. U.S. Department of Education. https://www.ed.gov/news/press-releases/statement-secretary-devos-2019-naep-results. Accessed 24 Feb 2021
7. Rimadana, M.R., Kusumawardani, S.S., Santosa, P.I., Erwianda, M.S.F.: Predicting student academic performance using machine learning and time management skill data. In: 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, pp. 511–515, December 2019. https://doi.org/10.1109/ISRITI48646.2019.9034585
8. bin Mohd Nasir, M.A.H., bin Asmuni, M.H., Salleh, N., Misra, S.: A review of student attendance system using near-field communication (NFC) technology. In: Gervasi, O., et al. (eds.) ICCSA 2015. LNCS, vol. 9158, pp. 738–749. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21410-8_56
9. Xu, X., Wang, J., Peng, H., Wu, R.: Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Comput. Hum. Behav. **98**, 166–173 (2019). https://doi.org/10.1016/j.chb.2019.04.015
10. Hasan, H.M.R., Rabby, A.S.A., Islam, M.T., Hossain, S.A.: Machine learning algorithm for student's performance prediction. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, pp. 1–7, July 2019. https://doi.org/10.1109/ICCCNT45670.2019.8944629
11. Scikit-learn: machine learning in Python—scikit-learn 0.24.2 documentation. https://scikit-learn.org/stable/. Accessed 04 May 2021
12. TensorFlow. https://www.tensorflow.org/. Accessed 04 May 2021
13. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/index.php. Accessed 28 Feb 2021
14. Nurafifah, M.S., Abdul-Rahman, S., Mutalib, S., Hamid, N.H.A., Malik, A.M.A.: Review on predicting students' graduation time using machine learning algorithms. Int. J. Mod. Educ. Comput. Sci. **11**(7), 1 (2019). https://doi.org/10.5815/ijmecs.2019.07.01

15. Lye, C.-T., Ng, L.-N., Hassan, M.D., Goh, W.-W., Law, C.-Y., Ismail, N.: Predicting pre-university student's mathematics achievement. Procedia. Soc. Behav. Sci. **8**, 299–306 (2010). https://doi.org/10.1016/j.sbspro.2010.12.041
16. Vijayalakshmi, V., Venkatachalapathy, K.: Comparison of predicting student's performance using machine learning algorithms. Int. J. Intell. Syst. Appl. **11**(12), 34 (2019). https://doi.org/10.5815/ijisa.2019.12.04
17. 3.1. Cross-validation: evaluating estimator performance—scikit-learn 0.24.2 documentation. https://scikit-learn.org/stable/modules/cross_validation.html. Accessed 04 May 2021