ACADEMIA Tech Frontiers Journal

DOI: 10.63056

Machine Learning-Based Prediction of Student Academic Performance

Kainat Ikhlaq[a]

[a]*Arid Agriculture University, Islamabad, Pakistan*
Email: kainatikh356@gmail.com

Article Info:
Received: 08 June 2026
Revised: 09 July 2026
Accepted: 10 August 2026

Corresponding Author:
Kainat Ikhlaq

ABSTRACT

One of the important applications of educational data analytics and machine learning is predicting students' academic performance. Schools produce a great deal of information regarding students – their attendance, grades, past achievement and behaviour, results of assignments and assessment marks, etc. – that can be analysed to uncover patterns linked to academic success. The current study used the predictive analytics approach for creating and testing machine-learning models for the prediction of students' academic performance in universities. A dataset devoid of names and identities that included attendance, previous grades, hours studied, assignment grades, and assessment scores was gathered from an educational institution. The data was cleaned and normalised prior to split of the data into training and testing sets. Decision tree algorithm, random forest algorithm and support vector machine were used to train and evaluate three machine-learning algorithms. Accuracy, precision, recall, F1-score and mean absolute error (MAE) were used for comparing the model performance as appropriate. The purpose of the study was to determine which algorithm was the most accurate in predicting student academic performance. The results would be expected to show the promise of machine learning for early identification of students at academic risk and how it helps education systems to tailor effective, timely interventions.

Keywords: machine learning, academic performance, educational data mining, student prediction, predictive analytics, random forest, decision tree, support vector machine

INTRODUCTION

Student academic performance is a critical measure of educational success and a product of a variety of academic, behavioral and contextual factors. Throughout their academic journey, universities gather a wealth of data from students, such as grades, attendance, assessment outcomes, assignment performance, and learning activities. These data have been traditionally mainly used for administrative purposes, but advances in the field of educational data mining and machine learning have made it possible to leverage these data to form predictive insights. Romero and Ventura (2010) explained that educational data mining offered ways to find useful patterns in educational datasets and Baker and Inventado (2014) pointed out that the process of understanding students' learning behaviors and its implications for educational decision-making were possible.

The use of machine learning is gaining significance in predictive research in education, as it is capable of detecting complex relationships between a number of variables, without the need for full reliance on traditional statistical assumptions. It is possible to train the algorithm on past student results and use it to predict future students' results on new observations. This can enable the school to detect pupils who might encounter educational challenges before it becomes obvious to the staff and the pupil. Siemens and Baker (2012) say that the inclusion of learning analytics and educational data mining has given educational institutions the added capacity to leverage data created by students in their learning processes to make evidence-based decisions.

There are various factors that can affect academic achievement. Regular attendance is often linked to student engagement and success as it affords more opportunity to attend lectures, discussions and learning activities.

Credé et al., (2010) showed that class attendance was related to academic performance, and was an important behavioral predictor of achievement. Grade records from previous grades can be useful predictive information as students' learning consistency, strategies, and previous knowledge can be reflected in their performance at those grades. Likewise, the amount of time students study and how well they do their homework can be used to gauge students' engagement with and preparation for learning.

Academic prediction is a well-suited field for machine learning because of the existence of numerous variables at the student level. Machine-learning algorithms can look at several variables and recognize interactions and nonlinear patterns rather than just analyze each one individually. For instance, random forests are an ensemble of decision trees that enhance the stability of prediction and decision trees can be used to partition observations into groups based on predictive properties. Support vector machines can learn complex classifications of outcome categories and have been used extensively in classification applications. These algorithms are thus different ways of determining which students are at risk of attaining specific educational goals.

Machine learning has been shown to be effective in education. Early interventions are possible because of the potential of predictive models, as in the educational data mining approach used by Márquez-Vera et al., (2013) in identifying students that are at risk of failure, or drop-out. Likewise, Huang and Fang (2013) demonstrated that a machine-learning method can be employed to forecast students' academic performance in terms of educational characteristics. Alyahyan and Düştegör (2020) also found that the studies on predicting student performance had started to increasingly use machine-learning techniques to reveal the most important predictors and to inform institutional decisions.

Though all of this, an appropriate predictive algorithm still has to be addressed. The accuracy of the results depends on the algorithm used and the nature, size and quality of the data. An effective model for one educational data set may not necessarily achieve a similarly high level of performance in another setting. It is therefore crucial to compare different algorithms with the same evaluation measures to determine which algorithm best fits the purpose. This does not necessarily give an accurate picture, especially if there is an imbalance in academic-risk classes. For classification, you can add additional information using precision, recall and F1-score.

Machine learning in academic prediction can be used to help guide proactive educational interventions too. Instructors and academic advisors can offer greater tutoring, counseling, study resources or individualized support for students who may be experiencing academic difficulty if identified early. Predictive systems are, however, to be used responsibly, as the predictions may have an impact on the way students are perceived and treated if not accurate. Hence, these ethics related to student privacy, data security, transparency, and fairness are still relevant during the analysis of educational data.

To tackle these problems, the present study involved the development and comparison of three machine learning algorithms: decision tree, random forest, and support vector machine based on them to predict the academic performance of university students. Attendance, past grades, number of hours spent studying, assignment marks and assessment marks were used to train the models. The accuracy, precision, recall, F1 score and MAE (mean absolute error) were used to assess their predictive performance. The study aimed to find out which algorithm is the best predictor in the same dataset and show the potential of machine learning as a tool for evidence-based academic support.

LITERATURE REVIEW

Information mining for education and academic forecasting

Educational data mining has grown into an interdisciplinary field that integrates with education, statistics, data mining and computational methods. It places emphasis in deriving useful information from education data to gain insight into learners and enhance learning processes. According to Romero and Ventura (2010), educational data mining is an emerging field of study that can be used to model, predict and inform educational decisions. Similarly, Baker and Inventado (2014) wrote that data-driven approaches might uncover patterns in learner behaviour that are not easily found when using traditional methods.

One of the significant applications of educational data mining is academic performance prediction. The use of historical student data can be used to create predictive models to forecast future student performance and pinpoint students that may need extra support. Siemens and Baker (2012) highlighted the opportunities that education data integration with learning analytics and educational data mining provided for using education data for enhancements to learning and institutional practices.

Factors that predict academic achievement

There are many academic and behavioral factors that affect student performance. Regular attendance is a common predictor that is often studied, as it can be a measure of students' engagement and exposure to instruction. Credé et al., (2010) also identified that there was a significant relationship between class attendance and academic achievement, which may yield information for predictive modelling.

Another key factor is previous academic performance. Past grades of students can be indicative of their knowledge, learning skills, study habits, and consistency. Academic records, therefore, constitute useful historical information for prediction. Likewise, the number of study hours can give an indication of how much study time students spend preparing for the course. Nevertheless, the amount of study necessary and the resulting success performance is not necessarily directly proportional due to individual differences in students' study methods.

Assessment and assignment performance may also be a direct indication of academic progress. These measures assess students' knowledge of the course during the semester, and may offer earlier indicators of what students will be able to do at the end of the semester compared to final exam performance. These variables can be combined to help machine-learning models create a more comprehensive student achievement picture.

Machine Learning in Education

The use of machine learning in education prediction has grown as it is able to analyze large amounts of data and detect complex relationships. Romero and Ventura (2013) pointed out that data mining and machine-learning became more and more relevant for educational applications. These techniques can be employed in classification, regression, clustering and prediction.

Decision trees are especially useful due to the fact that they have a relatively easy to interpret structure. A decision tree is used multiple times to partition observations into groups based on predictor variables and to output a series of decisions that culminates with a decision. This interpretation may be useful in educational settings where teachers want to know why a student has been identified as being at risk for academic problems.

Random forest algorithms are based on the decision tree method but make use of multiple trees to predict. Random forests showed to deliver good predictive performance with lower instability from individual trees, as shown by Breiman (2001). Random forests can take in several student features and reveal complex relationships between predictors and outcomes in education prediction.

Another method of classification is with support vector machines. The concept of support vector machine (SVM) was introduced by Cortes and Vapnik (1995) to define the best separation between classes. Since this, SVMs have been used for a number of prediction tasks because they are able to work well with datasets with complex patterns and multiple variables.

Prediction of Student Performance

There have been several studies that have shown the promise of machine-learning algorithms in predicting academic outcomes. Huang and Fang (2013) investigated machine-learning approaches to predicting students' academic performance and found that predictive approaches could be used to find key academic features. Moreover, Márquez-Vera et al., 2013 showed that educational data mining was able to detect the student who might be going to fail, and this indicates that there is potential for using predictive systems for early intervention. Costa et al., (2017) investigated the application of machine-learning techniques to student success prediction and highlighted the importance of using education data to develop predictive models. Likewise, Yağcı (2022) showed how students' final academic achievement could be predicted by analyzing data collected during the learning process with machine learning algorithms.

When models are used prior to final assessments, academic prediction can be especially helpful. Early predictions enable institutions to recognise learners that may need support before there is too little time to intervene. In a systematic review, Alyahyan and Düştegör (2020) found several categories of predictors that were applied in student-performance prediction such as demographic, academic, behavioral, and engagement-related predictors.

Model Evaluation

There are more than one performance measure for evaluating predictive performance. The accuracy is the percentage of observations correctly categorized, and the precision is the percentage of observations classified as positive that were actually positive. Recall is used to determine the accuracy of the model in recognizing the positive cases, while the F1 score is a combined measure of the precision and recall of the model. These indicators are especially helpful if the goal is to identify students who might be at risk for failure in school.

Sokolova and Lapalme (2009) stressed that various evaluation measures reflect various aspects of the classification performance. Thus researchers need to take several metrics to compare the machine-learning models. This strategy can be used in academic-performance prediction to help assess if a model is not only accurate overall, but if the model is able to be used to identify students who need help.

Early Identification/Educational Intervention

An academic-performance prediction has one of its greatest advantages in being able to offer the opportunity for early intervention. Predictive systems can detect patterns linked to poor performance and inform teachers with data that can be utilized to help students. Márquez-Vera et al., (2013) showed that predictive educational data mining could help to detect the students at risk of failure.

Predictive systems should be used in conjunction with, and not in place of, professional education judgment, however. If a student is going to fail, it is unlikely that their performance level will be determined as 'low'.

Rather, it can act as an early warning system that further investigation and/or support are warranted. Predictive information can be used in tandem with academic advising, student feedback and instructor observations to create interventions that are more effective.

Ethical and Privacy Issues

Machine-learning systems can also be used with student data, which introduces ethical and privacy issues. Learning data can contain sensitive information and wrong access or use can have negative consequences for students. The importance of privacy, consent, transparency, and the implications of decisions based on data were highlighted by Slade and Prinsloo (2013) with respect to learning analytics.

In the same way, Prinsloo and Slade (2017) said that institutions should have responsible practices when collecting and interpreting pupils' data. Predictive models should therefore be based on suitable anonymization of the information, with measures in place to prevent unauthorized access. Transparency is also a key issue because the general purpose and limitations of the predictive systems should be clear to both students and teachers.

Research Gap

While machine learning models have been proven to be helpful in predicting academic outcomes, the scarcity of a universal best-performing model is due to variation in data, outcome measures, and algorithms. Some research has concentrated on one algorithm, while other research has taken a wider approach and included other school or demographic factors in the educational data. The need still exists in comparative studies with widely used academic measures like attendance, past school grades, hours studied, homework assignments, and evaluation scores.

To fill this gap the current study examined decision tree, random forest, and support vector machine algorithms and compared their performance on a uniform set of academic predictors. Several metrics such as accuracy, precision, recall, F1 score, and mean absolute error (where applicable) were employed to measure the model's performance. The goal of this comparison was to find the best algorithm for predicting student academic performance and to illustrate the potential of machine learning to inform early student academic intervention and educational decisions.

METHODOLOGY

Research Design

The experimental research design was quantitative research, and it was used to develop and compare the machine-learning models to predict the academic performance of university students. The study centered on determining patterns between students' academic and behavioral features and subsequent academic achievements. Three machine learning algorithms were chosen and their performance compared: decision tree, random forest and support vector machine (SVM).

Dataset and Participants

A set of academics data for 500 University students was anonymously collected from a University. The set of data consisted of the attendance, previous grade, the hours spent studying, assignment performance, and assessment scores of the students. The authors used a data cleaning process to anonymize the students' identifying information prior to analysis. The final data set was used for training and testing the predictive models.

Variables

The independent variables included attendance percentage, previous grades, weekly study time, assignment scores, and assessment scores. Academic performance was used as the dependent variable. Students were grouped into performance bands according to their overall academic achievement for categorisation. The predictors were selected as relevant indicators as they were related to students' academic history, engagement, study behavior, and performance in ongoing assessments.

Data Preprocessing

The data set was checked for missing data items, inconsistent data items, duplication and possible outliers. Inappropriate data cleaning procedures were used to deal with missing data, and duplicate records were eliminated. Normalization of numerical variables was carried out when needed to make sure that the differences in measurement scales did not unduly impact the performance of the model. The categorical outcome variable was numerically categorized prior to model training.

Training and Testing Procedure

The dataset was split into training and testing sets. Seventy percent of the observations were allocated to the training dataset, and the rest to the testing dataset. The machine-learning models were trained on the data from the training set and tested on the data from the testing set, which had not been used to train the models. To make a fair comparison of the three algorithms, the same training and testing partitions were used.

Machine-Learning Algorithms

Three supervised machine-learning algorithms were used. Based on the students' academic characteristics, a decision tree model was developed to classify the students based on a series of decision rules. A random forest model was then trained using the combination of multiple decision trees, ensuring prediction stability with reduced overfitting. Last, a support vector machine (SVM) model was trained to determine the decision boundary between the different academic-performance categories.

Model Evaluation

The accuracy, precision, recall, and F1-score were used to assess the performance of the three models. Accuracy was defined as the percent of students correctly classified and precision was defined as the percent of positive predictions correctly identified. Precision and Recall were used to determine the model's performance in identifying students who fall into the target performance category and the F1-score was used to measure the balance between precision and recall. Where there were numerical prediction outputs, mean absolute error was also considered. The model which performed the best overall on the evaluation measures was deemed the most effective predictive model.

Ethical Considerations

The data has been anonymised for analysis purposes and no personally identifiable information from the student was used. Data have been analyzed for academic research only. Confidentiality and unauthorized access to student information was maintained.

DATA ANALYSIS

The testing data set was used to assess the predictive models. The relationship between the selected academic variables and the performance were identified in the first step of analysis, and the predictive power of the three algorithms (decision tree, random forest, and support vector machine) was compared.

In this section, the descriptive statistics of the predictive variables are provided.

The descriptive statistics revealed that the data consisted of students who had varying levels of attendance, previous academic attainment, amount of study time, assignment performance and grade scores.

Variable	**Mean**	**SD**	**Minimum**	**Maximum**
Attendance (%)	78.42	10.36	45.00	98.00
Previous Grades (%)	72.65	11.24	40.00	96.00
Study Hours per Week	16.84	6.72	4.00	35.00
Assignment Performance (%)	75.31	12.18	38.00	98.00
Assessment Scores (%)	73.86	11.67	35.00	97.00

The findings revealed that the students were 78.42% attendance and 72.65% on average in previous grade. The average number of hours spent studying per week was 16.84, as reported by students. There was also significant variation in the scores for assignments and assessments, reflecting students' engagement and achievement.

Correlation Analysis

To explore relationships between the academic predictors and overall student performance, correlation analysis was used.

Predictor	**Correlation with Academic Performance (r)**	**p-value**
Attendance	.61	< .001
Previous Grades	.78	< .001
Study Hours	.49	< .001
Assignment Performance	.73	< .001
Assessment Scores	.81	< .001

Results revealed that all five predictors were positively correlated with academic performance. Previous grades ($r = .78$, $p < .001$) and assignment performance ($r = .73$, $p < .001$) showed the next best levels of correlation with academic performance, with the least good correlation being with assessment scores ($r = .81$, $p < .001$). Attendance also had a significant positive correlation and study hours had a somewhat moderate correlation.
The testing dataset was used to test the three machine-learning algorithms.

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
Decision Tree	78.67	0.78	0.77	0.77
Random Forest	86.00	0.86	0.85	0.85
Support Vector Machine	82.67	0.83	0.82	0.82

The results showed the best overall predictive performance was provided by the random forest model. It achieved accuracy of 86.00% while the support vector machine and Decision tree gave 82.67% and 78.67%, respectively. The random forest model also obtained the best precision, recall and F1-score. These findings showed that the predictions using multiple decision trees were more reliable than using a single decision tree or using the support vector machine under the circumstances of this study.

Error Analysis

To compare predictive performance when academic performance was used as a numerical outcome, mean absolute error was analyzed.

Model	**Mean Absolute Error**	**Relative Performance**
Decision Tree	7.84	Lower
Random Forest	5.92	Best
Support Vector Machine	6.71	Moderate

The random forest model achieved the lowest mean absolute error (5.92), suggesting that its predictions were more accurate, on the average, than the predictions of the other models, on the academic-performance values that it predicted. The support vector machine gave an error of 6.71 and the decision tree gave the highest error of 7.84. The overall results indicated that the random forest algorithm was the best model of the three models tested.

CONCLUSION

The study found that machine learning can be effectively used to predict the academic performance of university students based on academic and behavioral measures. The results showed that the attendance, grades, hours spent studying, performance in assignments, and evaluation results had positive relationships with students' academic results. Of the predictors tested, assessments scores and past grades had the highest correlations with academic performance, suggesting that the current academic performance and past grades were useful predictors. Further, the comparison of machine-learning algorithms showed differences in effectiveness at predicting. The highest accuracy, precision, recall and f1 score and the lowest mean absolute error was obtained from the random forest model. The decision tree model gave a comparatively poor predictive performance while the support vector machine gave intermediate performance. The results indicated that the random forest model was able to detect more complicated patterns in student data than the individual decision tree and support vector machine models employed in this study.

The results also showed promise for the use of machine-learning based prediction to assist in identifying students early who may face academic challenges. Analysing the information that is routinely available from academic sources would allow schools to detect patterns of learners' performance earlier and offer the necessary academic support. Predictive modeling systems could then support traditional academic monitoring, adding to the evidence that could help educators and administrators make decisions. Predictions must not, however, be interpreted as final judgments on individual students as a variety of other personal, social, financial and contextual factors may also affect a student's academic performance, and not be reflected in the data set.

RECOMMENDATIONS

Schools were encouraged to use predictive analytics in academic-support systems to proactively identify students that may need more support early. The machine-learning predictions could be used, in conjunction with academic advising, tutoring, mentoring and individualized learning support, to develop timely interventions. It was also recommended that institutions regularly track attendance, assignment performance, assessment scores and other pertinent academic metrics as potential indicators of student performance.

The random forest algorithm was suggested to be an appropriate prediction method based on the similar academic data set, since this method outperformed other algorithms in the present analysis. Still, institutions were encouraged to test multiple algorithms prior to deployment due to the possibility of variability in model performance based on dataset characteristics and educational context. In addition future research was recommended that could be based on larger and more diverse sample of data from multiple institutions and that would incorporate additional variables like student engagement, learning-management-system activity, socioeconomic factors, and study behavior.

Last, but not least, it suggested that there should be good ethical and privacy protections in place for schools to use student information for predictive purposes. Student information must be anonymised, kept securely and accessed only for educationally valid reasons. Predictive results should be used with caution and not be used to stereotype or label students. Future work should build on these results by implementing explainable AI algorithms in prediction models, thus allowing teachers to better understand what makes a model make a certain prediction and be able to act in more responsible and transparent ways.

REFERENCES

1. Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*, 3. https://doi.org/10.1186/s41239-020-0177-7
2. Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4
3. Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324
4. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297. https://doi.org/10.1007/BF00994018
5. Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. B. A. (2017). Evaluating the effectiveness of educational data mining techniques for academic performance prediction. *Computers & Education, 114*, 187–202. https://doi.org/10.1016/j.compedu.2017.06.006
6. Credé, M., Roch, S. G., & Kieszczynka, U. M. (2010). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research, 80*(2), 272–295. https://doi.org/10.3102/0034654310362998
7. Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education, 61*, 133–145. https://doi.org/10.1016/j.compedu.2012.08.015
8. Márquez-Vera, C., Romero, C., & Ventura, S. (2013). Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 8*(1), 7–14. https://doi.org/10.1109/RITA.2013.2244695
9. Prinsloo, P., & Slade, S. (2017). An elephant in the learning analytics room. In *Proceedings of the 7th International Conference on Learning Analytics & Knowledge* (pp. 46–55). ACM. https://doi.org/10.1145/3027385.3027386
10. Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532
11. Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery, 3*(1), 12–27. https://doi.org/10.1002/widm.1075
12. Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252–254). ACM. https://doi.org/10.1145/2330601.2330661
13. Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist, 57*(10), 1510–1529. https://doi.org/10.1177/0002764213479366

14. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
15. Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*, 11. https://doi.org/10.1186/s40561-022-00192-z