Discover Education

Case Study

Performance prediction using educational data mining techniques: a comparative study

Yaosheng Lou[1] · **Kimberly F. Colvin**[1]

Received: 6 November 2024 / Accepted: 18 April 2025
Published online: 13 May 2025

Abstract
Predicting student performance has been a critical focus of educational research. With an effective predictive model, schools can identify potentially at-risk students and implement timely interventions to support student success. Recent developments in educational data mining (EDM) have introduced several machine learning techniques that can effectively analyze students' demographic information, learning processes, and other contextual factors to predict academic outcomes. However, limited research has compared the predictive accuracy of these EDM techniques with traditional statistical methods in real-world educational settings. This case study aims to address this gap by empirically evaluating the performance of generalized linear regression, decision tree, and random forest regression in predicting three end-of-course exams. The data are from a statewide high school dataset. Model performance was assessed using R-square, RMSE, MAE, and MSE. The results indicated that generalized linear regression consistently outperformed decision tree and random forest regression in terms of both predictive accuracy and error. Additionally, this study examined the capacity of these methods to identify important predictors. These findings may offer valuable insights for researchers and educators in selecting appropriate methods for similar prediction tasks.

Keywords Student performance · Educational data mining · Generalized linear regression · Random forest · Decision tree · Performance prediction · Predictive accuracy

Over the past decade, researchers have increasingly turned to big data techniques to extract valuable information from large datasets [1–4]. As a result, data mining has gained significant attention across various fields. Initially, data mining techniques were predominantly applied in business [5, 6] and scientific research [7, 8] due to the availability of large and structured datasets in these fields. There has been a growing interest in applying data mining techniques within educational settings, leading to the development of the field of educational data mining (EDM; [4]). The goal of EDM is to apply various data mining techniques to analyze and extract meaningful patterns from educational data, ultimately enhancing teaching, learning, and institutional decision-making. As Collier et al. [9] noted, EDM techniques are pivotal for conducting large-scale analyses in education. Indeed, educational databases are typically vast, containing records ranging from thousands to millions [10]. Although traditional statistical methods such as logistic regression [11] and generalized linear regression [12, 13] have been widely used to address educational problems, these approaches rely on predefined models (e.g., linear relationships) and specific assumptions about data distributions (e.g., normality). These limitations reduce their flexibility and computational efficiency when applied to complex and large-scale datasets. Consequently, the integration of data mining techniques into education has become essential to effectively address these challenges.

✉ Yaosheng Lou, ylou@albany.edu; Kimberly F. Colvin, kcolvin@albany.edu | [1]Department of Educational and Counseling Psychology, University at Albany-SUNY, 1400 Washington Ave, Albany, NY 12222, USA.

To begin, it is necessary to note that EDM has varied definitions in the literature. A widely accepted definition describes it as a field that applies computational techniques to educational datasets to improve understanding of learning processes, predict outcomes, and optimize educational systems [14]. In contrast, some researchers such as Baker and Yacef [15] define EDM as an interdisciplinary field that builds on statistical and computational foundations, suggesting that traditional statistical methods (e.g., multiple linear regression) should be also integrated into its broader analytical framework. For this paper, we adopt the first definition by Romero and Ventura [14], which does not consider traditional statistical methods as part of EDM techniques. In other words, only computational approaches are referred to as EDM techniques in this context.

While EDM is a relatively new field of research, it has experienced rapid growth. Ozyurt et al. [16] reviewed 2792 studies from 2008 to 2022, reporting a steady increase in publications from just 12 in 2008 to 359 in 2022. The authors identified 12 key topics[1] in the field of EDM. Among these topics, "performance prediction" was the fastest-growing topic. Many other researchers also acknowledge that "performance prediction" was the most researched and emerging topics in EDM in recent years [9, 17]. Specifically, student performance can be measured by various metrics, including quantitative outcomes (e.g., scores and GPA) and categorical outcomes (e.g., program completion, pass/fail in a course, and performance levels). Despite the growing popularity of EDM, it should be noted that research on its application to quantitative measures of student performance remains relatively limited compared to its use for categorical measures [17, 18].

Random forest regression and decision tree analyses are widely used to predict quantitative academic outcomes. A study by Doz and colleagues [19] serves as a practical example of random forest regression. The authors predicted students' grades on mathematics test using gender, geographical region, socioeconomic status, and students' origin. The results indicated that school type, region, and socioeconomic status are influential predictors, whereas gender and origin are less impact in prediction. Another study [20] used attendance, gender, high school type, high school score, delivery mode, and GPA to predict student course grades ranging from 0 to 100. GPA and high school score were identified as the most important predictors of a course grade. Al-Barrak and Al-Razgan [21] conducted a case study using decision trees to predict student GPA based on transcript grades. Their prediction model helped identify at-risk students in advance and highlighted the most critical courses in students' study plans.

When student performance is measured using categorical metrics (e.g., program completion, pass/fail in a course), various EDM methods are available, including decision tree, random forest classification, support vector machines, and clustering. For instance, Beaulac and Rosenthal [22] used random forest classification to predict whether students would obtain an undergraduate degree based on number of credits attempted, the mean grade in previous courses, the performance in first-year courses. Similarly, Asif et al. [23] used pre-university marks and examination marks of early years at university to predict student graduation performance. Their results revealed that decision trees could effectively perform this regression task and identify key predictors. Additionally, some studies have explored clustering [24–26], support vector machines [27], Naïve Bayes [28, 29] for predicting categorical student performance. With the application of these EDM techniques, educational institutions are now better able to predict and explain students' academic outcomes across all levels of education.

Although EDM techniques offer innovative approaches for predicting student performance, traditional statistical methods continue to dominate many educational prediction tasks. As Alyahyan and Düştegör [30] stated, current EDM techniques are mainly accessible only to educators with proficiency in artificial intelligence. Du et al. [17] reviewed 411 predictive studies published between 2000 and 2019 and found that traditional regression methods accounted for 54.25% of the techniques used for predicting student performance. Among these, generalized linear regression (GLR), a traditional statistical method, is widely used to predict quantitative outcomes, such as students' first year grades [12], GPA [13], and exam scores [31]. GLR has proven to be a more flexible and accurate method for prediction tasks compared to multiple linear regression, especially in contexts involving complex relationships [32, 33].

Given the strength of GLR, researchers have compared its performance with EDM techniques for prediction in many fields, including earth science [34–37], sociology [38], biomedicine [39]. Their results, however, showed discrepancies in different fields. For example, random forest was reported to have better performance than GLR in predicting fire hazard levels [35], soil fertility [37], and temperature prediction [36], in contrast, GLR was found to be superior to random forest in predicting species richness [40]. While researchers provide some insights that can help explain the discrepancy (e.g., GLR may be better suited for count data), there is yet no consensus that EDM techniques are superior to GLR with which type of dataset. More importantly, to our knowledge, no research has examined their comparative performance in educational contexts.

[1] Twelve key topics in EDM are: learning pattern and behavior, performance predictions, recommendation systems, sentiment and feedback analysis, online learning platforms, learning analytics, clustering student's profile, knowledge tracing, rule-based algorithms, feature selection, dropout risk prediction, and unstructured data analysis.

This study aims to fill this gap by evaluating the predictive performance of generalized linear regression, decision tree, and random forest using data from a K-12 educational setting. According to a recent systematic review by Ersozlu et al. [41], decision trees and random forests are the two most frequently used EDM techniques for educational data. Therefore, by comparing these methods, this case study can provide valuable insights into their comparative strengths and limitations, contributing to the deeper understanding of prediction methods in educational settings. The findings also offer practical guidance for researchers and educators in selecting appropriate methods for similar prediction tasks.

In this manuscript we will first describe the data, then define the three analytical techniques that will be compared: generalized linear regression, decision tree, and random forest. There is then a discussion of the four measures that will be used to compare the models generated by each technique. The detailed results of the models are provided by analytical techniques for each of the three predicted outcomes, then the results are compared across the three analytical techniques. Finally, practical guidance is offered and directions for future research are discussed.

1 Methods

1.1 Data description

The data for the current study came from the Massachusetts Department of Elementary and Secondary Education (https://www.doe.mass.edu/infoservices/research), which contained 17,957 high school students enrolled in 2019 academic year. A total of 13 variables with potential significance for performance prediction were included in this study, see Table 1. Among them, five were quantitative predictors, five were categorical predictors, and the remaining three served as outcomes. Note that each analysis targeted only one of these outcomes: current Math score, current English Language Arts (ELA) score, or current Science and Technology/Engineering (STE) score. In other words, three replicate analyses are included in this study to provide more generalized and comparable results.

1.2 Data preprocess

To enable effective analysis, cases with missing data were excluded, resulting in a final dataset of 16,454 cases. This approach is also known as complete case analysis [42]. This approach is generally not recommended because the remaining cases may form a biased subsample, leading to biased results in the analysis [43]. However, it is acceptable in the context of this comparative study, as the focus is on the predictive accuracy of the methods rather than the interpretation of specific results.

Table 1 The List of Variables

Variables	Description	Type	Role
County	County of residence	Categorical	P
School Type	Official school organizational type (Special Ed School/Collaborative; Public School; Charter School; Out-of-state School)	Categorical	P
EL	English learner status with two levels (Yes/No)	Categorical	P
EcoD	Economically disadvantaged status with two levels (Yes/No)	Categorical	P
Disability	Disability status with two levels (Yes/No)	Categorical	P
MASchool Years	Number of years student has attended MA schools	Continuous	P
School Years	Number of years student continuously enrolled in the same school	Continuous	P
District Years	Number of years student continuously enrolled in the same district	Continuous	P
Prior Math	Math score prior year	Continuous	P
Prior ELA	ELA score prior year	Continuous	P
Math Score	Current Math score	Continuous	O
ELA Score	Current English Language Arts (ELA) score	Continuous	O
STE Score	Current Science and Technology/Engineering (STE) score	Continuous	O

P = predictor; O = outcome

For the purpose of validation, all cases were randomly split into training (70%) and test (30%) datasets using RStudio. The training dataset was used to develop the models, while the test dataset was reserved for evaluating the performance of the fitted models. This approach allowed for an assessment of predictive accuracy by comparing the predicted values with the observed outcomes.

All analyses were conducted using the most current version of R (4.4.2) and RStudio (2024.09.1 + 394). The annotated R code for running these analyses can be obtained by contacting the first author of this article. This allows interested readers to extend this study in other contexts.

1.3 Generalized linear regression

Generalized linear regression (GLR) is a practical application of generalized linear models (GLMs), focusing on linear relationships between predictors and an outcome variable. In the present study, *glm* function in base R was used to conduct GLR analyses. Default parameters were used for Gaussian GLM with link = "identity", as the model assumes a direct and linear relationship between predictors and the outcome.

1.4 Decision tree

Decision tree (DT) is a machine learning method commonly used for predicting both categorical and quantitative outcomes. It operates by recursively splitting the dataset into subsets based on the feature that provides the maximum information gain or the best-fit split. The structure of a decision tree resembles a flowchart, as shown in Fig. 1. It begins with a root node which includes the entire samples. Internal nodes (e.g., B, C, D in Fig. 1) represent decisions based on certain features of the data (e.g., *a*1, *b*1). Ultimately, the tree extends to leaf nodes (such as *y*1, *y*2, …*y*6) which represent final predictions. In this study, decision tree analyses were conducted using the *rpart* package [44].

1.5 Random forest regression

Random forest regression (RFR) is a specific application of random forest algorithm designed for predicting quantitative outcomes [45]. This technique constructs multiple decision trees from random subsets of the data, with the final prediction being the average of the predictions from all trees. In this study, RFR analyses were conducted using the *randomForest* package [46]. The number of trees (*ntree*) was set to 1000, a commonly used value that balances accuracy, stability, and computational efficiency [47, 48]. Additionally, the number of features considered at each split (*mtry*) was set to three, which is approximately one-third of the total number of predictor variables, as recommended for regression tasks [49]. Bootstrap sampling (*bootstrap* = *True*) and the split quality criterion (*criterion* = *"squared_error"*) were left at their default settings, aligning with established practices for regression analyses [46].

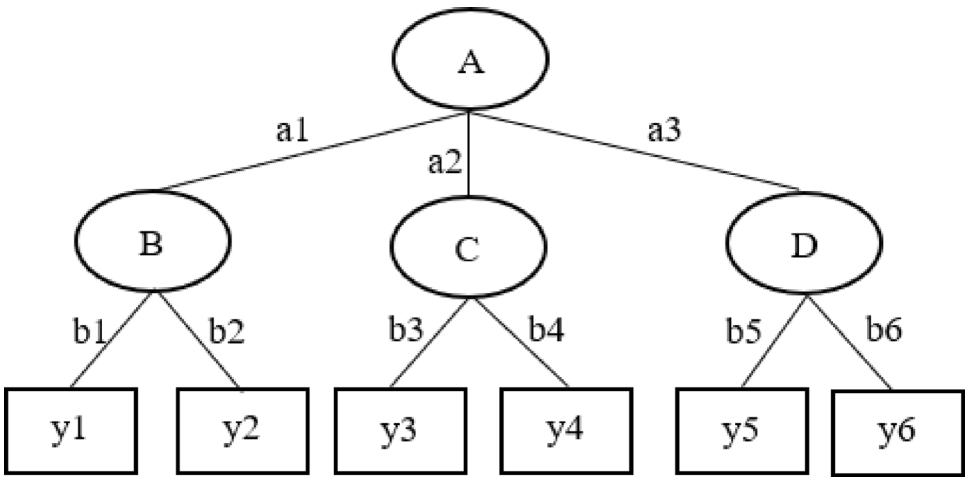


Fig. 1 Example of a Decision Tree Diagram

1.6 Models performance measures

The performance of the models was assessed using four metrics: R^2, root mean square error (RMSE), mean absolute error (MAE), and mean square error (MSE). These metrics are widely used in comparative studies [49–51]. The detailed descriptions of these metrics are discussed below.

1.7 R-square

R^2 Represents the proportion of the variance in the dependent variables explained by the independent variables in a regression model. Higher R^2 values indicate better predictive accuracy, with 1 representing a perfect fit.

1.8 Root mean square error

Root mean square error (RMSE) measures the magnitude of the prediction errors by calculating the square root of the average squared differences between predicted and observed values. RMSE is particularly valuable for assessing the overall accuracy of the model. Smaller RMSE values indicate better model fit.

1.9 Mean absolute error

Mean absolute error (MAE) measures the average magnitude of errors between predicted values and actual values ignoring their direction. MAE is essential for understanding the typical error magnitude in a model, especially when large outliers are less critical to the analysis. Smaller MAE values indicate better predictive performance.

1.10 Mean square error

Mean square error (MSE) measures the average squared difference between the actual values and the predicted values in regression tasks. It is closely related to RMSE but focuses directly on squared errors, making it particularly sensitive to large prediction errors. This sensitivity allows MSE to detect extreme outliers that may not be as evident with MAE or RMSE. A smaller MSE indicates greater predictive accuracy.

2 Results

2.1 Generalized linear regression

Before proceeding with the analyses using GLR, assumptions were checked to ensure the validity of the model (i.e., linearity, homoscedasticity and independence of errors). Failure to meet the assumptions may result in biased estimates and unreliable conclusions [52]. Since parameters obtained from a Gaussian GLR fit are similar to those from ordinary least square (OLS) regression (e.g., multiple linear regression; [53]), graphical residual plots were used to assess the model assumptions (see Fig. 2 as an example). No obvious abnormal pattern was identified in the plots of three GLR analyses. Specifically, the residuals versus fitted plot showed that the residuals were randomly scattered around the horizontal line at zero, with no discernible patterns. This indicates that the assumption of linearity was reasonably satisfied. The Q-Q plot further supported this, as the residuals closely followed the diagonal line. The residuals versus leverage plot revealed no significant patterns, and most residuals fell within Cook's distance boundaries, indicating the absence of highly influential points and supporting the independence of errors. Additionally, the scale-location plot and the residuals versus leverage plot did not indicate any evidence of heteroscedasticity (i.e., the variance of the residuals was constant). Overall, the diagnostic plots suggest that the assumptions of GLR analyses in this study were reasonably met. Additionally, tolerance and VIF measures were examined for multicollinearity issues, and no violations were detected.

Next, three GLR analyses were conducted to predict MATH, ELA, and STE separately. Table 2 summarizes the model performance based on R^2, RMSE, MAE, and MSE. The R^2 values ranged from 0.679 to 0.792. All metrics consistently indicated that GLR perform best when predicting MATH, with the highest R^2 value (0.792) and lowest error rates across RMSE (10.49), MAE (8.11), and MSE (110.04). On the other hand, the worst performance was observed when predicting

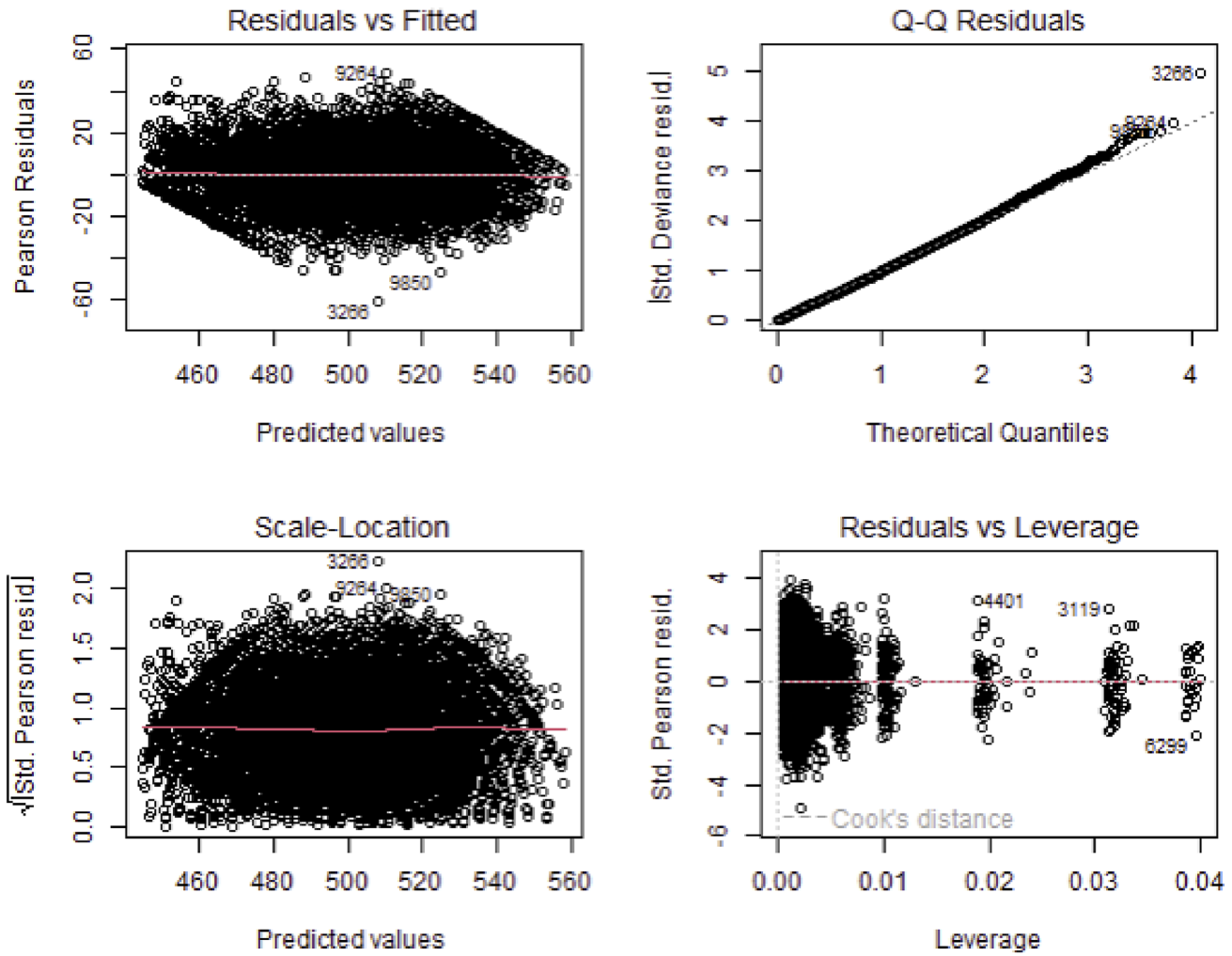


Fig. 2 The Diagnostic Plots When Predictor is ELA

STE, with the lowest R^2 value and highest error rates (i.e., RMSE = 12.91, MAE = 10.18, and MSE = 166.63). To enhance the interpretation of the model, the six most important predictors were identified based on statistical significance (*p*-value)

Table 2 Results of R^2, RMSE, MAE, and MSE in GLR Models

	MATH	ELA	STE
R^2	0.792	0.742	0.679
RMSE	10.49	12.31	12.91
MAE	8.11	9.69	10.18
MSE	110.04	151.66	166.63
Most important predictors (In order)	Prior Math	Prior ELA	Prior Math
	County = Middlesex	County = Norfolk	Prior ELA
	Prior ELA	Prior Math	County = Dukes
	County = Essex	School Type = Public School	County = Franklin
	County = Norfolk	School Type = Charter School	County = Suffolk
	County = Hampshire	Disability	EL

These predictors are not the only significant variables in the models

and ranked by their standardized beta coefficients, shown in Table 2. These predictors represent the key variables with the strongest influence on the model's outcomes.

2.2 Random forest regression

The performance of RFR models in predicting MATH, ELA, and STE are shown in Table 3. The R^2 values ranged from 0.669 to 0.783. Similar to the results of GLR models, RFR showed best performance when predicting MATH (i.e., $R^2 = 0.783$, RMSE = 10.67, MAE = 8.33, and MSE = 113.95) and worst performance when predicting STE (i.e., $R^2 = 0.669$, RMSE = 13.10, MAE = 10.33, and MSE = 171.71). Figure 3 shows the variable importance for each RFR model. The top six most important variables identified by RFR were listed in Table 3.

2.3 Decision tree

Table 4 summarizes the results of the decision tree models for predicting MATH, ELA, and STE. The R^2 values ranged from 0.621 for STE to 0.750 for MATH. Consistently, the decision tree model showed the best performance when predicting MATH and the poorest performance when predicting STE. The decision nodes used in the models are also listed in Table 4 (more information see Appendices A-C). Interestingly, prior Math emerged as the only critical predictor for both MATH and STE. For ELA, however, both prior Math and prior ELA were identified as key predictors.

2.4 Comparison of three methods

To compare the performance of the three different methods, four metrics were independently compared for predicting the same outcome, see Fig. 4. Across all prediction tasks, GLR consistently outperformed the other methods, showing the highest R^2 and the lowest RMSE, MAE, and MSE. For example, in predicting MATH, GLR achieved an R^2 of 0.792, RMSE of 10.49, MAE of 8.11, and MSE of 110.04, while the decision tree performed the worst with an R^2 of 0.750, RMSE of 11.47, MAE of 9.04, and MSE of 131.63. RFR showed intermediate performance, falling between GLR and the decision tree across all metrics.

The most important variables in predicting the target outcomes differed across the three methods. Overall, both the GLR and RFR models recognized "County" as a significant predictor, but only the GLR model specified which level of "County" contributed most to the predictions. When examining each method individually, more distinct patterns emerged. For MATH predictions, the GLR model identified several different counties as critical predictors, whereas the RFR model placed it behind "Disability". In contrast, the DT model relied solely on "Prior Math" as the significant predictor. For ELA predictions, the GLR model included "School Type" as a significant predictor, but the RFR model rated it as the least important factor (see Fig. 4). The DT model only included prior ELA and prior Math as key predictors. In the prediction of STE, the GLR model identified "English Learner" as an important variable, while the RFR model did not deem it important (see Fig. 4). Again, the DT model relied solely on "Prior Math" as the predictor for STE.

Table 3 Results of R^2, RMSE, MAE, and MSE in RFR Models

	MATH	ELA	STE
R^2	0.783	0.731	0.669
RMSE	10.67	12.58	13.10
MAE	8.33	9.91	10.33
MSE	113.95	158.23	171.71
Most important predictors (In order)	Prior Math	Prior ELA	Prior Math
	Prior ELA	Prior Math	Prior ELA
	Disability	Disability	County
	County	County	Disability
	EcoD	EcoD	EcoD
	School year	School years	School years

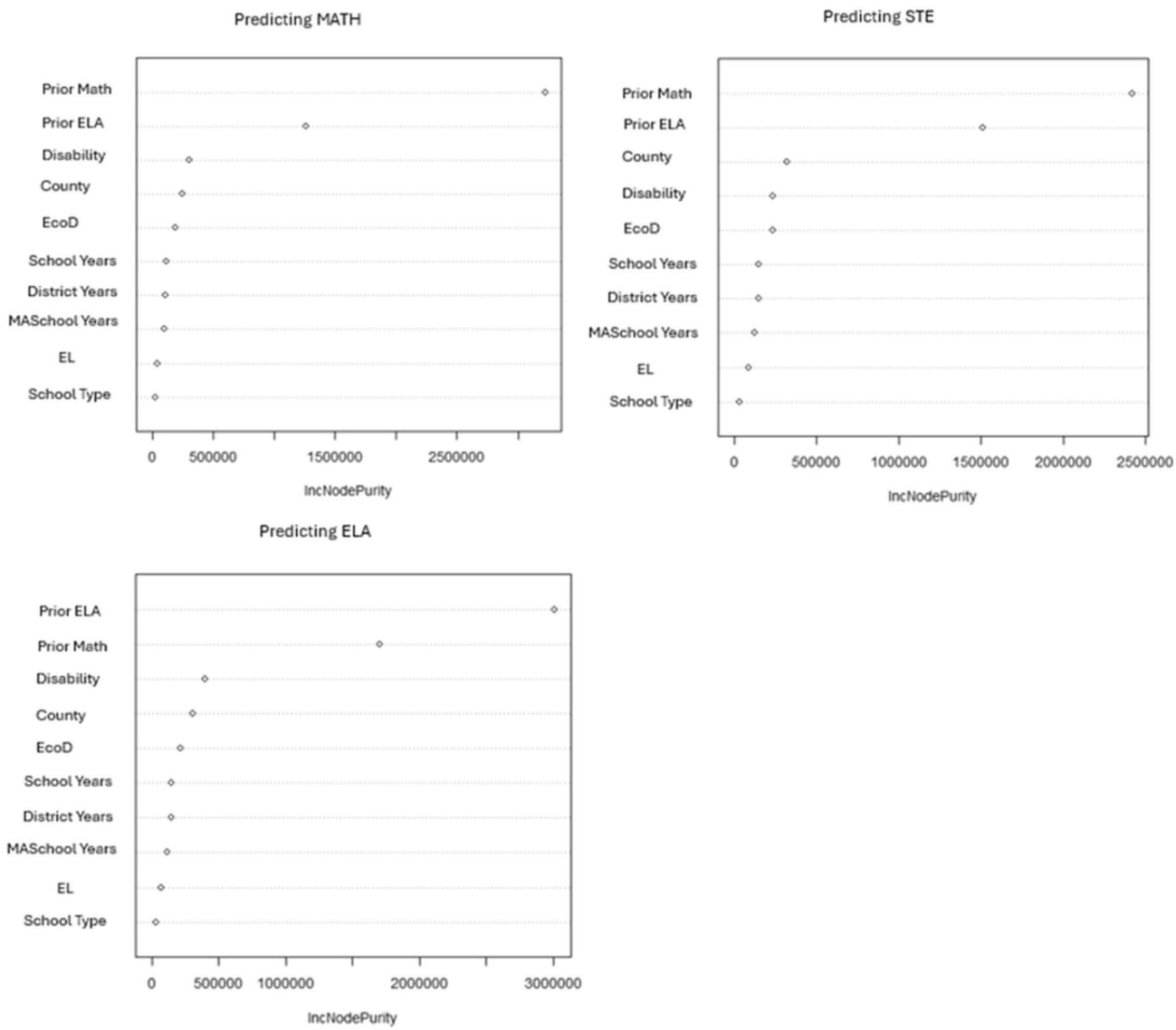


Fig. 3 Variable Importance for the Three RFR Models

Table 4 Results of R^2, RMSE, MAE, and MSE in DT Models

	MATH	ELA	STE
R^2	0.750	0.681	0.621
RMSE	11.47	13.69	14.03
MAE	9.04	10.81	11.16
MSE	131.63	187.39	196.72
Decision nodes (Unordered)	Prior Math	Prior ELA Prior Math	Prior Math

3 Discussion

In recent years, the comparison of machine learning techniques and traditional statistical methods has become an active area of research. However, comparative studies in the context of educational data remain limited compared to other fields, particularly in predicting quantitative student performance. Such studies are crucial for understanding

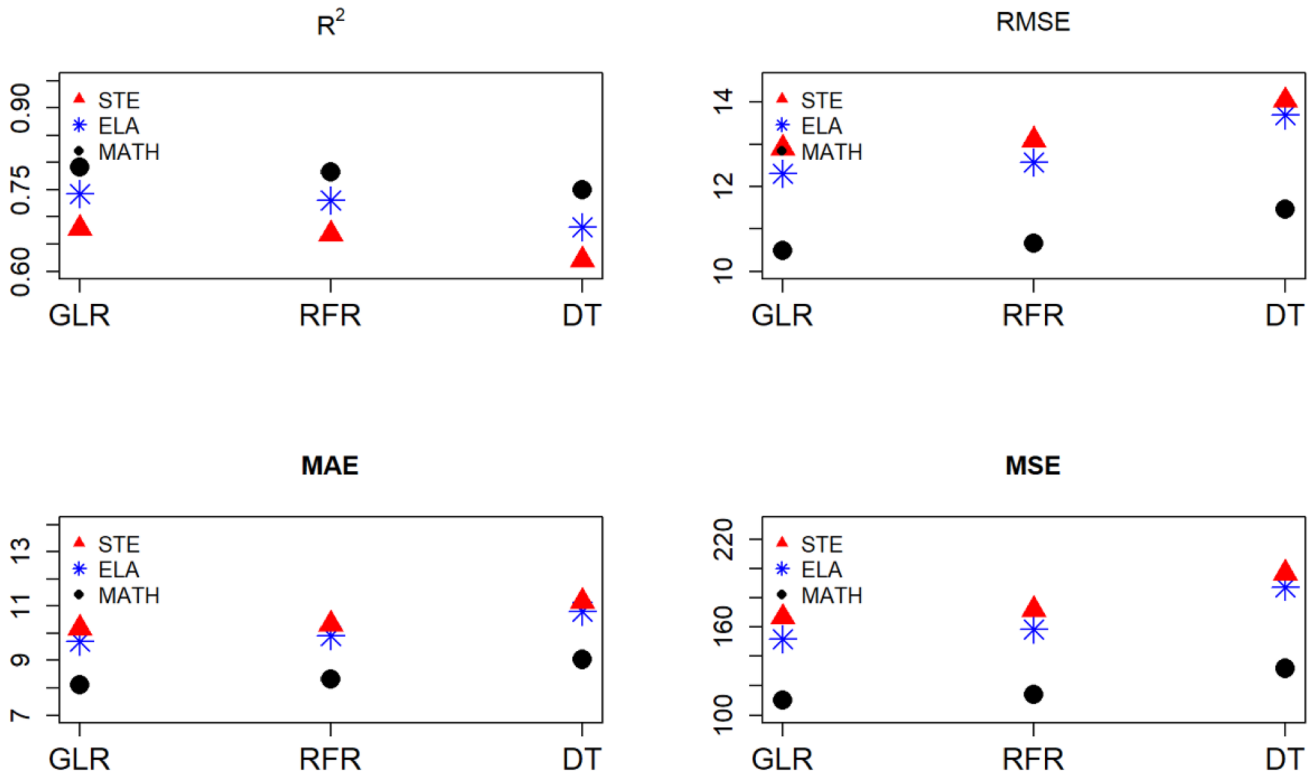


Fig. 4 Comparison of Model Performance Based on Each Metric. GLR = generalized linear regression, RFR = random forest regression, DT = decision tree

the strengths and limitations of these approaches in addressing the unique complexities of educational data. In particular, educational data often include categorical variables with multiple levels and academic records, which are critical attributes for making predictions. This study contributes to this discussion through a case study utilizing large-scale data from a K-12 educational setting. In the present study, the traditional methods of generalized linear regression and two machine learning algorithms in EDM techniques (i.e., random forest regression and decision tree) were examined and discussed.

According to three replicate analyses (i.e., predicting MATH, ELA, and STE), the GLR demonstrated advantages over RFR and DT for predicting educational outcomes in this dataset. This finding is not surprising. In prior research, GLR was also observed to outperform random forest in specific prediction tasks, such as predicting plant species richness [40]. The superior performance of GLR can be attributed to its use of coefficients and statistical tests to identify significant predictors, particularly when the outcome variable has a strong correlation with quantitative predictors. For example, in the prediction of MATH scores, prior MATH scores serve as a highly correlated predictor in the model. In contrast, RFR is often referred to as a "black box" model because its internal workings are complex and not easily interpretable. It evaluates the contribution of variables empirically through recursive partitioning across multiple decision trees. As Strobl et al. [54] observed in a simulation study, random forest can overemphasize the variables with many unique values, which can reduce model performance. As a result, RFR may be less suitable for datasets with variables like prior scores and county, as seen in this dataset. Notably, GLR also exhibited a significant advantage in processing speed, with each analysis processed in 1 s compared to the 164 s on average required by RFR. Lastly, DT, while effective for categorical predictions, has limitations when applied to quantitative variables. First, it is highly sensitive to noise in the training data [55, 56], making it prone to overfitting and increasing prediction errors. Moreover, even when predicting continuous outcomes, DT still divides the feature space into non-overlapping regions, assigning a constant prediction to all points within each region. For example, when predicting MATH scores, a DT might predict that all students with prior MATH scores above 528 will receive a final MATH score of exactly 529.2 (see Appendix A). This lack of continuity limits its ability to capture complex and continuous relationships in the data effectively.

However, we acknowledged that RFR remains a highly promising technique in predicting quantitative outcomes, particularly for datasets where the assumptions of GLR are violated. Many studies have examined the ability of RFR in

capturing complex patterns in multidimensional or nonlinear datasets [57, 58]. Moreover, while DT exhibited worst performance in predicting any of three academic outcomes, its simplicity and ease of interpretation are strengths that researchers frequently highlight [59, 60]. This tradeoff between accuracy and interpretability suggests that researchers may opt for DT models when their practical applications prioritize interpretability over predictive precision.

Furthermore, the important predictors identified by GLR and RFR were different. This discrepancy may arise from the differences in their methodological frameworks. GLR assumes a linear relationship between predictors and the outcome. Variable importance in GLR models can be assessed both by *p*-value and standardized beta coefficient. However, RFR functions as a "black box" model. Variable importance in RFR is often evaluated using metrics such as the mean decrease in impurity (i.e., Gini importance, which is a measure of variable importance) or permutation importance. These metrics evaluate the contribution of each predictor to reducing uncertainty or improving prediction accuracy across trees. On the other hand, GLR provides the advantage of evaluating the effect of each specific level of a categorical variable on the outcome, whereas RFR evaluates only the overall importance of the variable. At this point, GLR may be a more suitable choice for educational researchers seeking to identify important predictors when working with categorical variables that have several levels.

This study is not without any limitations. First, this comparative study was conducted using a specific dataset, limiting the generalizability of the findings to other educational datasets. However, when the data characteristics and outcomes are similar enough to the ones included in this case study, then the findings may offer a direct guide on method selection. Second, the assessment of variable importance in GLR relied on *p*-values and standardized beta coefficients, which is not the only available approach. Grömping [48] summarized various metrics suited to different data characteristics for identifying important variables in regression models. Future comparative studies could explore additional metrics to evaluate the performance of GLR in identifying key variables. Lastly, the present study focused solely on predicting a quantitative academic outcome. A relevant direction for future research is to examine whether EDM techniques outperform traditional statistical methods in predicting categorical academic outcomes (e.g., pass/fail in a course) within specific educational datasets. In such cases, more traditional statistical methods could be evaluated alongside EDM techniques to gain deeper insights into their relative performance across different educational contexts. These comparisons would provide practical guidance for practitioners in selecting the most effective methods for prediction tasks.

Author contributions Y.L. contributed to the literature search, study design, data analysis, and manuscript's writing. K.C. contributed to the manuscript's writing, supervision, and validation. All authors reviewed and approved the final manuscript.

Funding No funding was received for conducting this study.

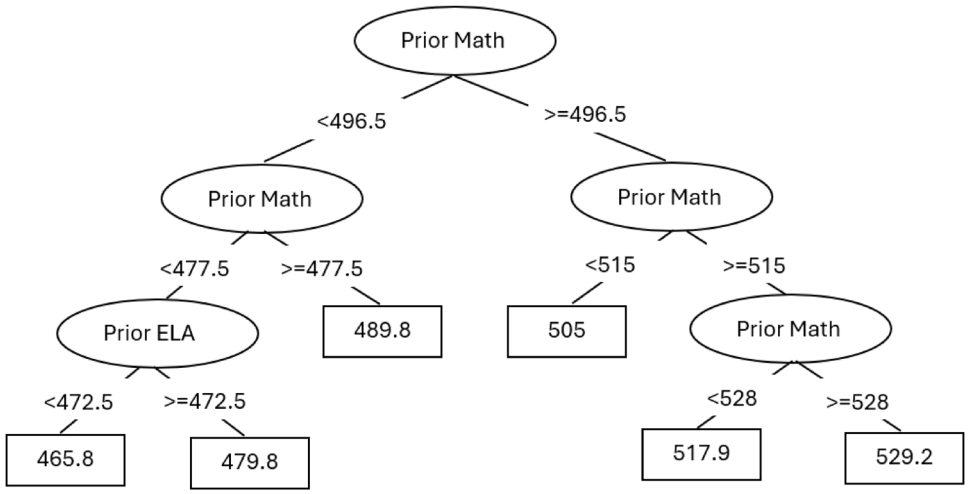
Data availability Data for this study is publicly accessible and can be obtained from the Massachusetts Department of Elementary and Secondary Education (https://www.doe.mass.edu/infoservices/research).

Declarations

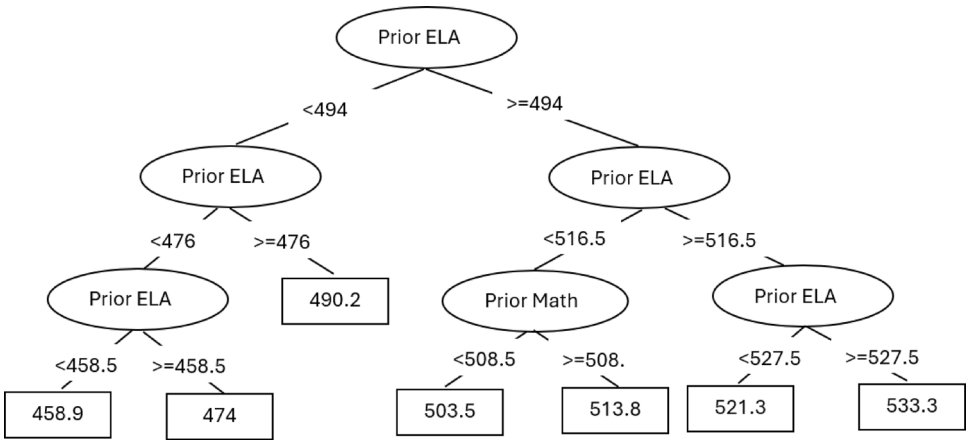
Ethics approval and consent to participate Ethics, Consent to Participate, and Consent to Publish declarations are not applicable to this study, as it used a publicly available secondary dataset. The authors were not involved in the data collection process. The dataset can be accessed through the Massachusetts Department of Elementary and Secondary Education (https://www.doe.mass.edu/infoservices/research).

Competing interests The authors declare no competing interests.

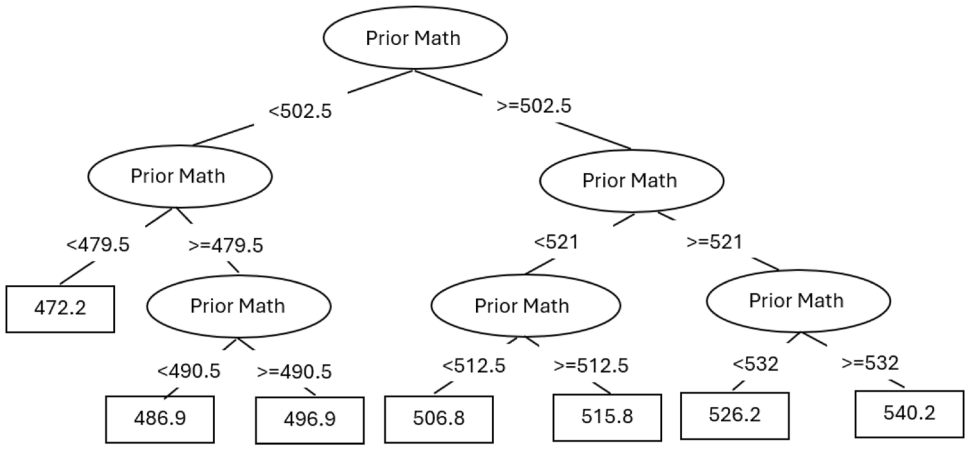
Appendix A: Decision tree diagram when predicting STE



Appendix B: Decision tree diagram when predicting ELA



Appendix C: Decision tree diagram when predicting MATH



References

1. Chen G, Rolim V, Mello RF, Gašević D. (2020). Let's shine together! A comparative study between learning analytics and educational data mining. Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, 544–553
2. Dutt A, Ismail MA, Herawan T. A systematic review on educational data mining. Ieee Access. 2017;5:15991–6005.
3. Nahar K, Shova BI, Ria T, Rashid HB, Islam AS. Mining educational data to predict students performance: a comparative study of data mining techniques. Educ Inf Technol. 2021;26(5):6051–67.
4. Xiao W, Ji P, Hu J. A survey on educational data mining methods used for predicting students' performance. Eng Rep. 2022;4(5):12482.
5. Bose I, Mahapatra RK. Business data mining—a machine learning perspective. Inf Manage. 2001;39(3):211–25.
6. Giudici P. Applied data mining: statistical methods for business and industry. John Wiley & Sons; 2005.
7. Chen SY, Liu X. The contribution of data mining to information science. J Inf Sci. 2004;30(6):550–8.
8. Kamath C. On mining scientific datasets. In: Kamath C, editor. Data mining for scientific and engineering applications. Boston: Springer; 2001. p. 1–21.
9. Collier Z, Sukumar J, Barmaki R. Discovering Educational Data Mining: An Introduction. Pract Assess Res Evaluat. 2024. https://doi.org/10.1007/978-1-4615-1733-7_1.
10. Khasanah AU. A comparative study to predict student's performance using educational data mining techniques. IOP Conf Series. 2017;215(1): 012036.
11. Costa SF, Diniz MM. Application of logistic regression to predict the failure of students in subjects of a mathematics undergraduate course. Educ Inf Technol. 2022;27(9):12381–97.
12. Anderton R, Chivers P. Predicting academic success of health science students for first year anatomy and physiology. Int J Higher Educ. 2016. https://doi.org/10.5430/ijhe.v5n1p250.
13. Arifin M, Widowati W, Farikhin F, Gudnanto G. A regression model and a combination of academic and non-academic features to predict student academic performance. TEM J. 2023;12(2):855.
14. 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.

15. Baker RS, Yacef K. The state of educational data mining in 2009: a review and future visions. J Educ Data Mining. 2009;1(1):3–17.
16. Ozyurt O, Ozyurt H, Mishra D. Uncovering the educational data mining landscape and future perspective: a comprehensive analysis. Ieee Access. 2023;11:120192–208.
17. Du X, Yang J, Hung JL, Shelton B. Educational data mining: a systematic review of research and emerging trends. Inf Discov Deliv. 2020;48(4):225–36.
18. Dol SM, Jawandhiya PM. Systematic review and analysis of EDM for predicting the academic performance of students. J Ins Eng. 2024;105(4):1021–71.
19. Doz D, Cotič M, Felda D. Random forest regression in predicting students' achievements and fuzzy grades. Mathematics. 2023;11(19):4129.
20. Nachouki M, Mohamed EA, Mehdi R, Abou Naaj M. Student course grade prediction using the random forest algorithm: analysis of predictors' importance. Trends Neurosci Educ. 2023;33: 100214.
21. Al-Barrak MA, Al-Razgan M. Predicting students final GPA using decision trees: a case study. Int J Inf Educ Technol. 2016;6(7):528.
22. Beaulac C, Rosenthal JS. Predicting university students' academic success and major using random forests. Res High Educ. 2019;60:1048–64.
23. Asif R, Hina S, Haque SI. Predicting student academic performance using data mining methods. Int J Comput Sci Netw Secur. 2017;17(5):187–91.
24. Bharara S, Sabitha S, Bansal A. Application of learning analytics using clustering data mining for students' disposition analysis. Educ Inf Technol. 2018;23:957–84.
25. Priyambada SA, Mahendrawathi ER, Yahya BN. Curriculum assessment of higher educational institution using aggregate profile clustering. Proc Comput Sci. 2017;124:264–73.
26. Alfiani AP, Wulandari FA. Mapping student's performance based on data mining approach (a case study). Agric Agric Sci Proc. 2015;3:173–7.
27. Buenaño-Fernández D, Luján-Mora S, Villegas-Ch W. 2017. Improvement of massive open online courses by text mining of students' emails: a case study. In Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality, 1–7.
28. Salinas JGM, Stephens CR. 2015. Applying data mining techniques to identify success factors in students enrolled in distance learning: A case study. In Advances in Artificial Intelligence and Its Applications: 14th Mexican International Conference on Artificial Intelligence, MICAI 2015, Cuernavaca, Morelos, Mexico, October 25–31, 2015, Proceedings, Part II 14, 208–219. Springer International Publishing.
29. Chen X, Vorvoreanu M, Madhavan K. Mining social media data for understanding students' learning experiences. IEEE Trans Learn Technol. 2014;7(3):246–59.
30. Alyahyan E, Düştegör D. Predicting academic success in higher education: literature review and best practices. Int J Educ Technol High Educ. 2020;17(1):3.
31. Nakayama M, Mutsuura K, Yamamoto H. 2018. Contributions of student's assessment of reflections on the prediction of learning performance. In 2018 17th International Conference on Information Technology Based Higher Education and Training (ITHET) (pp. 1–4). IEEE.
32. Fox J. Applied regression analysis and generalized linear models. London: Sage publications; 2015.
33. McCullagh P. Generalized linear models. New York: Routledge; 2019.
34. Chiaverini L, Macdonald DW, Hearn AJ, Kaszta Ż, Ash E, Bothwell HM, Cushman SA. Not seeing the forest for the trees: Generalised linear model out-performs random forest in species distribution modelling for Southeast Asian felids. Ecol Inf. 2023;75:102026.
35. Eskandari S, Amiri M, Sādhasivam N, Pourghasemi HR. Comparison of new individual and hybrid machine learning algorithms for modeling and mapping fire hazard: a supplementary analysis of fire hazard in different counties of Golestan province in Iran. Nat Hazards. 2020;104:305–27.
36. Nordin ND, Zan MSD, Abdullah F. 2020. Comparative analysis on the deployment of machine learning algorithms in the distributed brillouin optical time domain analysis (BOTDA) Fiber sensor. In Photonics (Vol. 7, No. 4, p. 79). MDPI.
37. Salmanpour A, Jamshidi M, Fatehi S, Ghanbarpouri M, Mirzavand J. Assessment of macronutrients status using digital soil mapping techniques: a case study in Maru'ak area in Lorestan province, Iran. Environ Monitor Assess. 2023;195(4):513.
38. Gill TM. 2019. Comparing random forest with generalized linear regression: predicting conflict events in western Africa (Master's thesis, The University of Arizona).
39. Song L, Langfelder P, Horvath S. Random generalized linear model: a highly accurate and interpretable ensemble predictor. BMC Bioinf. 2013;14:1–22.
40. Lopatin J, Dolos K, Hernández HJ, Galleguillos M, Fassnacht F. Comparing generalized linear models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile. Remote Sens Environ. 2016;173:200–10.
41. Ersozlu Z, Taheri S, Koch I. A review of machine learning methods used for educational data. Educ Inf Technol. 2024;29:1–21.
42. Pigott TD. A review of methods for missing data. Educ Res Eval. 2001;7(4):353–83.
43. Schlomer GL, Bauman S, Card NA. Best practices for missing data management in counseling psychology. J Couns Psychol. 2010;57(1):1.
44. Therneau T, Atkinson B. rpart: recursive partitioning and regression trees. R package version. 2023;4(1):23.
45. Schonlau M. Random forests. In: Schonlau M, editor. Applied statistical learning: with case studies in stata. Cham: Springer International Publishing; 2023. p. 183–204.
46. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18–22.
47. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
48. Grömping U. Variable importance assessment in regression: linear regression versus random forest. Am Stat. 2009;63(4):308–19.
49. Kushwah JS, Kumar A, Patel S, Soni R, Gawande A, Gupta S. Comparative study of regressor and classifier with decision tree using modern tools. Mater Today. 2022;56:3571–6.
50. Smith PF, Ganesh S, Liu P. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. J Neurosci Methods. 2013;220(1):85–91.
51. Xie X, Wu T, Zhu M, Jiang G, Xu Y, Wang X, Pu L. Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land. Ecol Ind. 2021;120: 106925.
52. Breslow NE. Generalized linear models: checking assumptions and strengthening conclusions. Statistica applicata. 1996;8(1):23–41.
53. Dobson AJ, Barnett AG. An introduction to generalized linear models. Boca Raton: Chapman and Hall/CRC; 2018.
54. Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2006, June). Bias in random forest variable importance measures. In *Workshop on statistical modelling of complex systems*. Citeseer.

55. Zhan H, Liu Y, Xia Y. 2022. Consistency of oblique decision tree and its boosting and random forest. arXiv preprint arXiv:2211.12653.
56. Lee S, Bikash KC, Choeh JY. Comparing performance of ensemble methods in predicting movie box office revenue. Heliyon. 2020;6(6):e04260.
57. Evans JS, Murphy MA, Holden ZA, Cushman SA. 2010. Modeling species distribution and change using random forest. In Predictive species and habitat modeling in landscape ecology: Concepts and applications (pp. 139–159). New York, NY: Springer New York.
58. Fife DA, D'Onofrio J. Common, uncommon, and novel applications of random forest in psychological research. Behav Res Methods. 2023;55(5):2447–66.
59. Jo N, Aghaei S, Benson J, Gomez A, Vayanos P. 2023. Learning optimal fair decision trees: Trade-offs between interpretability, fairness, and accuracy. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 181–192.
60. Mienye ID, Jere N. 2024. A survey of decision trees: concepts, algorithms, and applications. IEEE Access.