Journal for STEM Education Research (2023) 6:358–374
https://doi.org/10.1007/s41979-023-00099-5

RESEARCH BRIEF

Predicting STEM Major Choice: a Machine Learning Classification and Regression Tree Approach

Chi-Ning Chang[1] · Shuqiong Lin[2] · Oi-Man Kwok[3] · Guan Kung Saw[4]

Accepted: 25 April 2023 / Published online: 11 May 2023

Abstract
Despite the increasing demand for professionals in science, technology, engineering, and mathematics (STEM), only a small portion of young people in the USA pursue a postsecondary degree in STEM. To identify the major predictors of STEM participation, this study uses a machine learning approach, a Classification and Regression Tree (CART), to analyze a wide range of individual, family, and school factors obtained from national survey data of US high school freshmen in fall 2009 who eventually enrolled in STEM college majors by 2016. The analytic results indicate that calculus credits, science identity, total STEM credits, and math achievement are the most predictive factors during the high school years of college STEM major selection. The CART-based tree also shows how these four variables interactively predict the likelihood of students enrolling in STEM college majors.

Keywords STEM college major · Course-taking · Motivation · Classification and regression tree · High School Longitudinal Study of 2009–2016

Despite the increasing demand for mid- and high-skilled professionals in science, technology, engineering, and mathematics (STEM), only a small portion of young people in the USA pursue a postsecondary degree in STEM (National Science Foundation, 2019; U.S. Bureau of Labor Statistics, 2017; U.S. Department of Education, 2015). Prior studies employing expectancy-value theory and social cognitive career theory have documented a variety of individual and contextual factors—such as mathematics achievement, science

✉ Chi-Ning Chang
changc10@vcu.edu

1 Virginia Commonwealth University, Richmond, VA, USA

2 American Institutes for Research, Austin, TX, USA

3 Texas A&M University, College Station, TX, USA

4 Claremont Graduate University, Claremont, CA, USA

self-efficacy, and financial support—that are related to STEM college major choice and career aspirations (Mau & Li, 2018; Wang, 2013; Wille et al., 2020). Less clear is *which factor plays a relatively more significant role* for adolescents who choose a STEM college major in pursuit of a STEM-related career. Identifying the most predictive factor(s) of STEM college major choice among high school students has important implications for efforts to increase STEM participation. This study addresses that critical research gap by analyzing the US nationally representative High School Longitudinal Study of 2009–2016 (HSLS:09–16) (National Center for Education Statistics [NCES], 2018) to identify individual, family, and school factors in adolescence that are *most predictive* of students entering a postsecondary STEM degree program.

The HSLS:09–16 study began with more than 23,000 US ninth graders (high school freshmen), their parents, math and science teachers, school administrators, and school counselors in fall 2009 (NCES, 2018). It collected a broad range of STEM-related variables, including both intrinsic factors (such as math interest and science identity) and extrinsic/behavioral factors (such as STEM course-taking and afterschool program participation) (NCES, 2018). We employed the Classification and Regression Tree (CART) algorithm, which uses a machine learning approach that permits auto-selection and furnishes the results with a tree structure, to help visualize how STEM-related variables influence students' decision-making related to STEM major choice (Steinberg & Colla, 2009a, b). This study is one of the first to apply the CART method to uncover the most predictive factors that influence pursuit of STEM degrees based on hundreds of variables in a nationally representative, longitudinal study.

Literature Review

Increasing opportunities for learners to choose STEM careers are a national priority (National Science Foundation, 2020). Given that most STEM workers (72.3%) have a college degree in STEM (U.S. Census Bureau, 2019), investigating the factors that influence a high school student's choice to pursue a STEM college major can help address this priority. Previous studies have identified various pre-college factors that might be associated with STEM college major choice. Students' demographic and family backgrounds are major factors. Specifically, female students, racial and ethnic minorities, and economically disadvantaged students tend to show lower interest in pursuing careers in STEM (Riegle-Crumb & Morton, 2017; Saw et al., 2018). Parents' occupations and involvement also influence students' STEM learning and career development (Howard et al., 2019; Moakler & Kim, 2014). Other factors include students' performance and motivation in STEM (Eccles, 1983, 2009; Lent et al., 1994; Saw & Chang, 2018; Wang, 2013), as well as their learning experiences and context factors in high school, such as school location (Saw & Agger, 2021), teacher quality (Althauser, 2015; Lee et al., 2015; Park et al., 2019), extracurricular opportunities (Kitchen et al., 2018; Franco & Patel, 2017; Means et al., 2016), and STEM course-taking (Gottfried & Bozick, 2016).

Although previous studies have collectively identified a broad range of factors that could potentially affect STEM college major choice, each study has only covered limited aspects due to the scope of the research. Some studies suggest that future research should include more potential exogenous variables, such as science-related motivational factors rather than merely math-related expectancy value constructs, to investigate the links between these factors and the pursuit of STEM career pathways (Gottfried & Bozick, 2016; Wang, 2013; Wille et al., 2020). In practice, all of these identified factors work simultaneously throughout the STEM career development process. Therefore, it is important not only to investigate what factors can predict the choice of a STEM college major, but also to identify how some of the factors can play a *relatively more significant role* than others in predicting the choice of college major.

To fill this literature gap, the present study includes a wider range of predictors collected by the HSLS:09–16 study. For students' demographics, we included predictors such as socioeconomic status (SES), gender, and race/ethnicity, as previous studies have shown that female students, racial and ethnic minorities, and low-income students are less likely to pursue STEM careers (Riegle-Crumb & Morton, 2017; Saw et al., 2018). For students' family backgrounds and parental involvement, we selected predictors such as parents' occupations and their support for math and science homework, as well as in-school and out-of-school STEM activities. These parental factors could benefit students' learning and career development in STEM by providing them with greater exposure and opportunities (Howard et al., 2019; Moakler & Kim, 2014). For students' career aspirations, motivation, and performance in STEM, we selected variables including their career and education goals, math and science self-efficacy, utility, identity, interest, and cost, as well as a range of performance measures (e.g., math standardized scores, GPA, SAT, ACT, AP and IB scores in STEM), based on expectancy-value theory, social cognitive career theory, and prior research (Eccles, 1983, 2009; Lent et al., 1994; Saw & Chang, 2018; Wang, 2013).

For teacher quality, we included unobserved factors that are critical to students' STEM learning achievement, such as math and science teachers' perceptions of professional learning communities, self-efficacy, expectations, collective responsibility, and principal support (Althauser, 2015; Lee et al., 2015; Park et al., 2019). For school location, we selected urbanicity and geographic region as predictors, given the geographic disparities in postsecondary STEM participation (Saw & Agger, 2021). For extracurricular opportunities, we included variables such as whether a school offers STEM-related programs (e.g., supporting underrepresented students in STEM and informing parents about college majors and careers in STEM), which may benefit students pursuing careers in STEM (Kitchen et al., 2018; Franco & Patel, 2017; Means et al., 2016). Since high school STEM course completion is positively linked to college major choice (Gottfried & Bozick, 2016), we included a list of STEM courses taken as predictors.

No prior studies have included such a large number of relevant variables to explore factors that could predict the choice of a STEM college major in high school students. The CART algorithm is a powerful tool for identifying factors with the most predictive power while unveiling how the selected factors interactively predict the STEM college major choice. This has never been applied in prior studies due to substantially smaller numbers of variables along with traditional analytic approaches (e.g., logistic regression in Lee's (2015) study, multilevel logistic regression in Bottia and colleagues' (2017) study, and Wang's (2013) study with the use of structural equation modeling). By examining a wide range of potential predictors and applying this advanced technique, this study could provide educators and policymakers with new perspectives and insights into which factors could be relatively more important in predicting the choice of a STEM college major among high school students.

Methods

Sample and Measures

The eligible sample from HSLS:09–16 is composed of 11,560 US high school students who participated in the 2009 base year, 2012 first follow-up survey, 2013–2014 updates and high school transcripts collection, and then reported their college majors in the 2016s follow-up survey. About 23% of these students majored in STEM. Guided by prior studies, we selected a wide range of 102 variables, including individual, family, and school factors. These variables are used simultaneously to predict students' college majors as either STEM or non-STEM. The list of variables for this study is provided in the Appendix.

Analytic Strategy

We employed the CART algorithm, implemented using the R package *rpart* (Therneau & Atkinson, 1997), to capture the complex mechanism of students' decision-making with regard to enrolling in STEM college majors by identifying a set of factors and explaining how those factors predict the students' decisions about enrolling in STEM majors. The algorithm was chosen due to its desirable properties: (a) it does not require strong model assumptions, which are typically needed when using traditional regression models; (b) it automatically identifies the important predictors and their linear/nonlinear relationships with outcomes (Lee et al., 2010; Steinberg & Colla, 2009a, b; Timofeev, 2004); (c) it is able to handle missing data without extra imputation procedures (Deconinck et al., 2005; Feelders, 1999; Verbyla, 1987); and (d) it is an interpretation-friendly algorithm compared with other "black-box" data-mining techniques.

First, to build a CART-based tree, the Gini index (Breiman et al., 1984; Steinberg & Colla, 2009a, b) was used to automatically select the important independent variables. The maximum depth of the tree was set at 30. Cost complexity (Breiman et al., 1984), with complexity parameters equaling 0.1, was chosen in the pruning process. Surrogate splitting (Feelders, 1999) was used to handle missing data for independent variables. Through these settings, the algorithm produced a pruned tree to predict the probability that a given student will declare a college major in STEM based on the selected predictors.

Second, to avoid model overfitting issues and to be able to evaluate predictive accuracy, the sample was split into training and testing datasets using the 80/20 rule (Anis et al., 2015; Zheng, 2004). Specifically, we used the random sampling method without replacement to select 80% of the samples (N_{train} = 9248) as the training data for developing the CART-based tree. The remaining 20% of the samples (N_{test} = 2312), who were not exposed to the tree development, served as the testing data to evaluate the predictive accuracy of the tree. In other words, we established the statistical model used to predict the outcome using the training dataset, and we used the testing data to validate the prediction through the established model. The measure of prediction accuracy was examined. A sensitivity analysis using random forest analysis was conducted to evaluate the consistency of the CART results. The CART algorithm also applied the student longitudinal analytic weight provided by HSLS:09–16. Hence, the results are weighted to represent US ninth graders in fall 2009.

Results

Figure 1 shows the output of the final CART-based tree, which predicts the probability of a student declaring a STEM college major. Out of all the independent variables, only four variables are deemed relatively more important and are automatically selected to construct the final tree: credits earned in calculus during high school, science identity in grade 11, total STEM credits earned during high school, and math achievement in grade 11. Therefore, these four variables play *relatively important roles* in a student's decision to choose a STEM college major.

With the final four predictors, the trained samples were split into five groups. As illustrated in Fig. 1, Group 1 is students (accounting for 81% of the high school students) who did not earn any credits in calculus and have a low

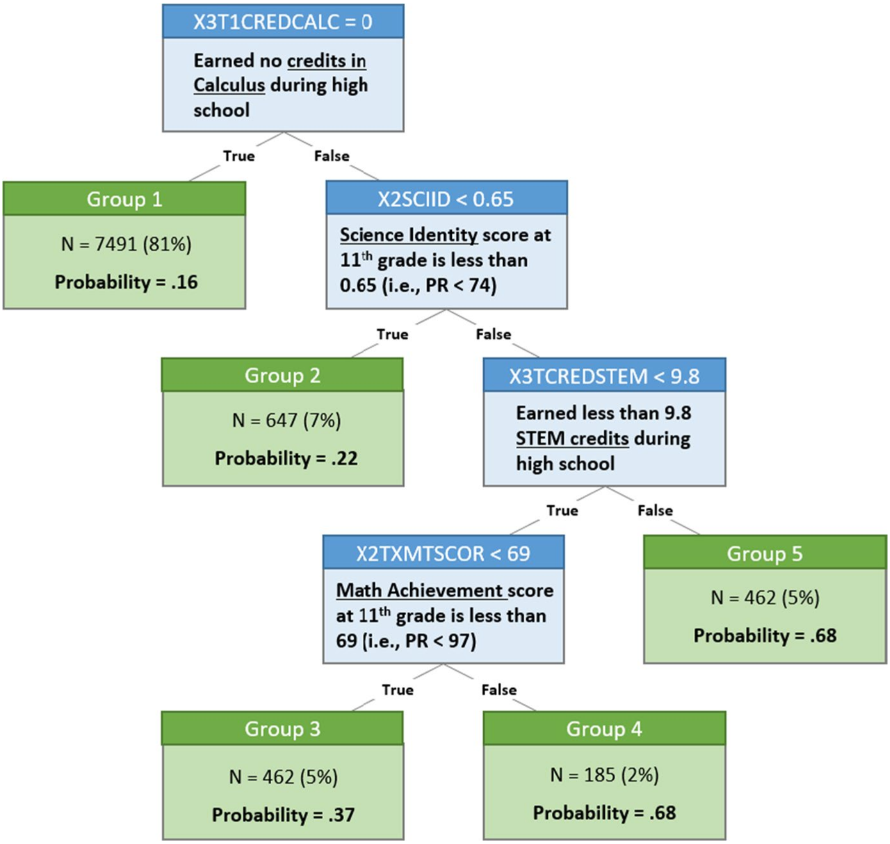

Fig. 1 The final CART-based tree. Probability indicates the chance of actually majoring in STEM. The HSLS:09–16 provides *Z* scores for science identity (X2SCIID) and *T* scores for math achievement (X2TXMTSCOR). To make these two standardized scores more comprehensible when interpreting the results, while also keeping the interpretation consistent across these two measures, we present the percentile ranks (PR) for these two measures converted from *Z* score and *T* score

probability of majoring in STEM (prob. = 0.16). Group 2 is students (accounting for 7% of the high school students) who earned credits in calculus and had a percentile rank (PR) for science identity in 11th grade < 74, and also have a low probability of selecting a college major in STEM (prob. = 0.22). Group 3 is students (accounting for 5% of the high school students) who earned credits in calculus, had a PR for science identity in 11th grade ≥ 74, earned fewer than 9.8 STEM credits during high school, and had a PR for math achievement scores in 11th grade < 97, and have a probability of declaring a STEM college major (prob. = 0.37).

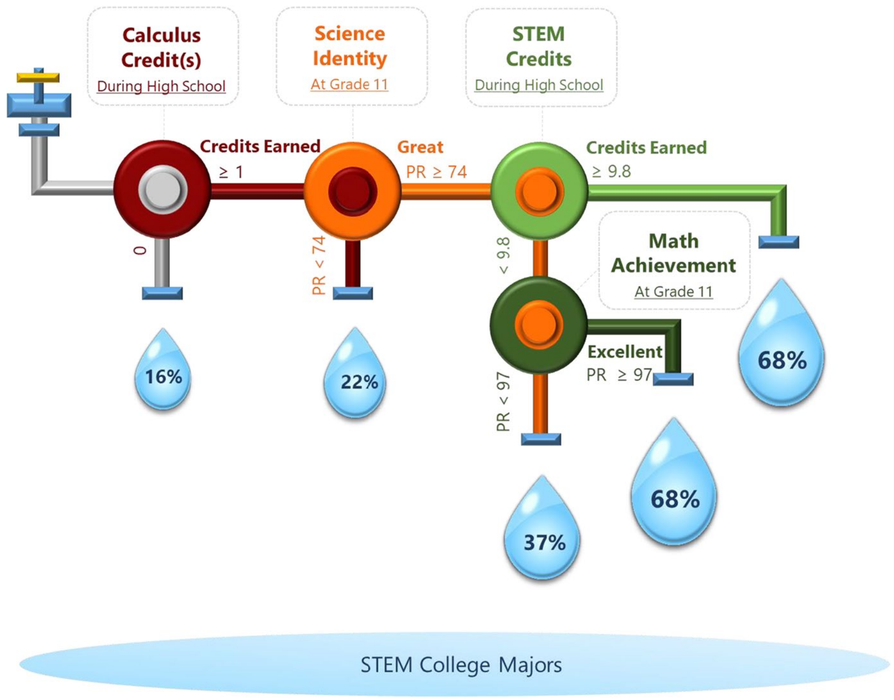

Fig. 2 STEM pipelines — predicting STEM college major choice in high school (2009–2016). The water drops represent the probability of declaring a college major in STEM. PR stands for percentile rank. The CART results illustrate how these four variables influenced the 2009 cohort's college major choice in 2016. In summary, if high school students do not earn any calculus credits, the likelihood of majoring in STEM disciplines will be only 16%. Furthermore, even if students earn calculus credit(s), their chance of pursuing a postsecondary STEM degree will still be low (22%) if they do not exhibit a high science identity in the 11th grade (PR < 74). On the other hand, if students earn calculus credit(s) and have a high level of science identity in the 11th grade (PR ≥ 74), the likelihood of enrolling in a STEM college major will increase substantially (from 16 to 37%). Interestingly, the probability of students majoring in STEM will be boosted to 68% if students earn calculus credit(s), have a high level of science identity in the 11th grade, and either earn at least 9.8 credits in STEM-related courses or have high math achievement in the11th grade (PR ≥ 97)

The remaining two groups have average probabilities larger than 0.5. Group 4 is students (accounting for 2% of the high school students) who had credits in calculus, a PR of science identity in 11th grade ≥ 74, total STEM credits < 9.8, and a PR for math achievement scores in 11th grade ≥ 97, and show the highest probability of enrolling in a STEM major in college (prob. = 0.68). Group 5 is students (accounting for 5% of the high school students) who earned credits in calculus, had a PR for science identity in 11th grade ≥ 74, and earned at least 9.8 credits in STEM (i.e., X3TCREDSTEM ≥ 9.8), and also have the highest probability of selecting a college major in STEM (prob. = 0.68).

Using the CART algorithm for prediction based on the test dataset (i.e., 20% of the full data) led to a classification accuracy equaling 0.80. A sensitivity analysis using random forest analysis indicated that the four selected variables in the CART are also identified as important variables using the mean decrease accuracy method, which further strengthened our confidence in the CART results. The CART-based tree can also be converted into Fig. 2, a more understandable image demonstrating how these four variables interactively predict STEM college major choice.

Discussion

Consistent with prior studies (Gottfried & Bozick, 2016; Riegle-Crumb et al., 2012), our findings suggest that completing at least one calculus class during high school is highly predictive of entering STEM fields. More importantly, our CART analysis is the first to demonstrate that calculus course completion is the most predictive factor among 102 examined variables, including individual, family, and school factors. Specifically, the probability of selecting a STEM college major is only 16% for students who do not earn any calculus credits during high school. This set of findings underscores the importance of offering and supporting the completion of advanced math coursework for high school students, particularly the study of calculus. Alarmingly, only about 50% of high schools in the USA offer calculus (U.S. Department of Education, 2016). Although our study does not include the school-level course offering variables, we could still speculate that students from the calculus-excluded schools might be more likely to have a lower rate of STEM participation.

Our study also uncovers that science identity is the second most predictive variable for enrolling in a postsecondary STEM degree program. It is important to note that science identity is relatively more significant when compared to other STEM motivational factors, including math self-efficacy, science interest, and STEM career aspiration. Science identity reflects how students act to convince themselves and others that they are science students, which is a powerful source of persistence in science (Robinson et al., 2019; Stets et al., 2017). Our CART study indicates that if students earn calculus credit(s) and report a high level of science identity (PR ≥ 74), the likelihood of choosing a STEM college major will increase substantially, to 37% and higher. School administrators and policymakers might consider developing or adopting programs or curricula that can help students cultivate a science identity in high school or at an earlier stage.

Predictably, students who earn more credits in STEM-related courses and have excellent math achievement in high school are more likely to enroll in STEM majors in college. However, these two factors (3rd and 4th most predictive variables) are "conditional" on the first two. In other words, only if students earn 9.8 or more STEM credits or demonstrate excellent math achievement, in combination with earning calculus credit(s) and having a high level of science identity, will the probability of declaring a STEM college major increase from 37 to 68%. This "conditional" implication, uncovered by the CART method, is a novel finding and addition to the literature on STEM education and career development.

There are four limitations to our study. First, our study relies on public-use secondary data. Other important predictors that are not released (e.g., school-level course offerings) or collected (e.g., neighborhood STEM resources) might be omitted. For example, the inclusion of school- and district-level data could provide insight into how related policies, resources, and programs contribute to students' STEM learning and pursuit of STEM careers. Due to this limitation, we are unable to determine the relative importance of the four selected variables compared to the variables omitted from this study. Although we could not include all possible predictors in our model, our study covers a broader range of aspects for prediction than previous studies. Second, we restrict the initial sample (23,000 + ninth graders from 944 schools in 2009) to those students ($n = 11,560$) who participated in the follow-up surveys from 2009 to 2016. We acknowledge that attrition bias might be a threat to internal and external validity. Therefore, to reduce the threat, our analysis has applied the student longitudinal analytic weight provided by the NCES. Third, our study results could only represent the findings for the 2009 ninth-grade cohort. Nonetheless, this cohort is the latest nationally representative, longitudinal high school sample for STEM education research conducted by the NCES. Fourth, some of our measures (e.g., parental involvement and STEM motivational factors) involve items with repeated measures. However, each of these items is measured only twice (grades 9 and 11) in high school. Due to the limited number of repeated measures, we use the regular CART for this study. Future longitudinal studies with repeated measures at multiple time points could consider employing the promising longitudinal CART algorithm (Kundu & Harezlak, 2019).

Despite these limitations, the findings of this study contribute to the current STEM literature in the following important ways: (a) identifying the relatively important variables among a rich set of predictors associated with STEM college major choice, (b) presenting how these four most predictive variables interactively predict the likelihood of choosing a STEM college major, and (c) demonstrating the potential of using the CART algorithm to uncover previously unexamined nuances of STEM educational and career pathways. Well-developed and effectively implemented programs could increase STEM participation and motivation (Hudson et al., 2020; Pike & Robbins, 2019). Our findings provide educators and policymakers with new perspectives and insights on which relatively important factors could be intervened among young students.

Appendix 1

Table 1 List of variables for the study

Variable name	Description	Data type
Dependent variable		
X4RFDGMJSTEM	X4 If first or second/double major is STEM	Categorical
Independent variables		
Demographics		
X1SES	X1 Student SES (composite score)	Continuous
X1SEX	X1 Student Gender	Categorical
X1RACE	X1 Student Race/Ethnicity	Categorical
Family backgrounds		
X1DADOCC_STEM1	X1 Father/male guardian's current/most recent occupation: STEM code 1 (sub-domain)	Categorical
X1MOMOCC_STEM1	X1 Mother/female guardian's current/most recent occupation: STEM code 1 (sub-domain)	Categorical
X2DADOCC_STEM1	X2 Father/male guardian's current/most recent occupation: STEM code 1 (sub-domain)	Categorical
X2MOMOCC_STEM1	X2 Mother/female guardian's current/most recent occupation: STEM code 1 (sub-domain)	Categorical
Parental involvement		
P1MTHHWEFF	P1 Confidence in helping with 9th-grade math homework	Categorical
P1SCIHWEFF	P1 Confidence in helping with 9th-grade science homework	Categorical
P1CAMPMS	P1 Participated in math or science camp outside of school in last year	Categorical
P1STEMDISC	P1 Discussed STEM program or article with 9th grader in last year	Categorical
P2MTHHWEFF	P2 Confidence in helping with math homework 2011–2012/when last enrolled	Categorical
P2SCIHWEFF	P2 Confidence in helping with science homework 2011–2012/when last enrolled	Categorical
P2STEMDISC	P2 Discussed STEM program or article with teenager in last year	Categorical
STEM career aspiration		
X1STU30OCC_STEM1	X1 Student occupation at age 30: STEM code 1 (sub-domain)	Categorical
X2STU30OCC_STEM1	X2 Student occupation at age 30: STEM code 1 (sub-domain)	Categorical

Table 1 (continued)

Variable name	Description	Data type
X1STUEDEXPCT	X1 How far in school 9th grader thinks he/she will get	Categorical
STEM motivation		
X1MTHEFF	X1 Math self-efficacy (composite score)	Continuous
X1MTHUTI	X1 Math utility (composite score)	Continuous
X1MTHID	X1 Math identity (composite score)	Continuous
X1MTHINT	X1 Math interest (composite score)	Continuous
X1SCIEFF	X1 Science self-efficacy (composite score)	Continuous
X1SCIUTI	X1 Science utility (composite score)	Continuous
X1SCIID	X1 Science identity (composite score)	Continuous
X1SCIINT	X1 Science interest (composite score)	Continuous
X2STUEDEXPCT	X2 How far in school 9th grader thinks he/she will get	Categorical
X2MTHEFF	X2 Math self-efficacy (composite score)	Continuous
X2MTHUTI	X2 Math utility (composite score)	Continuous
X2MTHID	X2 Math identity (composite score)	Continuous
X2MTHINT	X2 Math interest (composite score)	Continuous
X2SCIEFF	X2 Science self-efficacy (composite score)	Continuous
X2SCIUTI	X2 Science utility (composite score)	Continuous
X2SCIID	X2 Science identity (composite score)	Continuous
X2SCIINT	X2 Science interest (composite score)	Continuous
X2BEHAVEIN	X2 Scale of school motivation (composite score)	Continuous
X2MEFFORT	X2 Scale of math class effort (composite score)	Continuous
X2SEFFORT	X2 Scale of science class effort (composite score)	Continuous
S1TEFRNDS	S1 Time/effort in math/science means not enough time with friends	Categorical
S1TEACTIV	S1 Time/effort in math/science means not enough time for extracurriculars	Categorical

Table 1 (continued)

S1TEPOPULAR	S1 Time/effort in math/science means 9th grader won't be popular	Categorical
S1TEMAKEFUN	S1 Time/effort in math/science means people will make fun of 9th grader	Categorical
Student academic performance and preparation		
X1TXMTSCOR	X1 Mathematics standardized theta score	Continuous
X1SCHOOLBEL	X1 Scale of student's sense of school belonging (composite score)	Continuous
X1SCHOOLENG	X1 Scale of student's school engagement (composite score)	Continuous
X2TXMTSCOR	X2 Mathematics standardized theta score	Continuous
X2EVERDROP	X2 Ever drop out	Categorical
X2PROBLEM	X2 Scale of problems at high school (composite score)	Continuous
X3EVERDROP	X3 Ever drop out	Categorical
X3TGPAMAT	X3 GPA: mathematics	Continuous
X3TGPAHIMTH	X3 GPA: highest level mathematics course taken	Continuous
X3TGPASCI	X3 GPA: science	Continuous
X3TGPAHISCI	X3 GPA: highest level science course taken	Continuous
X3TGPAENGIN	X3 GPA: engineering/engineering tech	Continuous
X3TGPASTEM	X3 GPA: STEM courses	Continuous
X3TGPATOT	X3 Overall GPA computed	Continuous
X3TGPAMTHAP	X3 GPA: AP/IB math courses	Continuous
X3TGPASCIAP	X3 GPA: AP/IB science courses	Continuous
S1HRMHOMEWK	S1 Hours spent on math homework/studying on typical school day	Categorical
S1HRSHOMEWK	S1 Hours spent on science homework/studying on typical school day	Categorical
C2AVGSATMATH	C2 Average SAT mathematics score	Continuous
C2AVGACTMATH	C2 Average ACT mathematics score	Continuous
C2AVGACTSCI	C2 Average ACT science score	Continuous
Math and science teacher quality		

Table 1 (continued)

X1TMCOMM	X1 Scale of math teacher's perceptions of math professional learning community (composite score)	Continuous
X1TMEFF	X1 Scale of math teacher's self-efficacy (composite score)	Continuous
X1TMEXP	X1 Scale of math teacher's perceptions of math teachers' expectations (composite score)	Continuous
X1TMPRINC	X1 Scale of math teacher's perceptions of principal support (composite score)	Continuous
X1TMRESP	X1 Scale of math teacher's perceptions of collective responsibility (composite score)	Continuous
X1TSCOMM	X1 Scale of science teacher's perceptions of science professional learning community (composite score)	Continuous
X1TSEFF	X1 Scale of science teacher's self-efficacy (composite score)	Continuous
X1TSEXP	X1 Scale of science teacher's perceptions of science teachers' expectations (composite score)	Continuous
X1TSPRINC	X1 Scale of science teacher's perceptions of principal support (composite score)	Continuous
X1TSRESP	X1 Scale of science teacher's perceptions of collective responsibility (composite score)	Continuous
School location		
X1LOCALE	X1 School locale (urbanicity)	Categorical
X1REGION	X1 School geographic region	Categorical
X2LOCALE	X2 School locale (urbanicity)	Categorical
X2REGION	X2 School geographic region	Categorical
Extracurricular programs		
C1PURSUE	C1 School has program to encourage underrepresented student in math/science	Categorical
C1INFORM	C1 School has program to inform parent about math/science higher ed/careers	Categorical
C1ENCCLG	C1 School has program to encourage students not considering college to do so	Categorical
C2ENCSTEM	C2 School has program to encourage underrepresented student in STEM	Categorical
C2INFSTEM	C2 School has program to inform parent about STEM higher ed/careers	Categorical
C2ENCCLG	C2 School has program to encourage students not considering college to do so	Categorical
STEM course-taking		
X3T1CREDALG1	X3 At least one credit earned in: algebra 1	Categorical
X3T1CREDALG2	X3 At least one credit earned in: algebra 2	Categorical
X3T1CREDINTM	X3 At least one credit earned in: integrated math	Categorical

Table 1 (continued)

X3T1CREDPREC	X3 At least one credit earned in: analysis/pre-calculus	Categorical
X3TCREDAPMTH	X3 Credits earned in: AP/IB mathematics courses	Continuous
X3T1CREDCALC	X3 At least one credit earned in: calculus	Categorical
X3T1CREDGEO	X3 At least one credit earned in: geometry	Categorical
X3T1CREDSTAT	X3 At least one credit earned in: statistics/probability	Categorical
X3T1CREDTRIG	X3 At least one credit earned in: trigonometry	Categorical
X3TCREDMAT	X3 Credits earned in: mathematics	Continuous
X3TCREDAPSCI	X3 Credits earned in: AP/IB science courses	Continuous
X3T1CREDBIOL	X3 At least one credit earned in: biology	Categorical
X3T1CREDCHEM	X3 At least one credit earned in: chemistry	Categorical
X3T1CREDESCI	X3 At least one credit earned in: geology/earth science	Categorical
X3T1CREDPHYS	X3 At least one credit earned in: physics	Categorical
X3TCREDSCI	X3 Credits earned in: science	Continuous
X3TCREDENGIN	X3 Credits earned in: engineering/engineering tech	Continuous
X3TCREDSTEM	X3 Credits earned in: STEM	Continuous
X3TCREDAPIB	X3 Credits earned in: AP/IB combined	Continuous
X3TCREDMTSC	X3 Credits earned in: combined mathematics and science	Continuous
Analytic weight		
W4W1STU	Student longitudinal analytic weight	Continuous

A search for the variable name in the codebook (http://nces.ed.gov/onlinecodebook) will reveal more details on value labels, variable description, frequency, and percentage *X1* composite variables/fall term of the 9th grade (2009), *X2* composite variables/spring term of the 11th grade (2012), *X3* composite variables/beyond high school graduation (2013–2014). *X4* composite variables/second follow-up (2016), *S1* student survey/fall term of the 9th grade (2009), *C1* counselor survey/fall term of the 9th grade (2009), *C2* counselor survey/spring term of the 11th grade (2012), *P1* parent survey/fall term of the 9th grade (2009), *P2* parent survey/spring term of the 11th grade (2012)

References

Althauser, K. (2015). Job-embedded professional development: Its impact on teacher self-efficacy and student performance. *Teacher Development, 19*(2), 210–225. https://doi.org/10.1080/13664530.2015.1011346

Anis, M., Ali, M., & Yadav, A. (2015). A comparative study of decision tree algorithms for class imbalanced learning in credit card fraud detection. *International Journal of Economics, Commerce and Management, 3*(12), 86–102.

Bottia, M. C., Mickelson, R. A., Giersch, J., Stearns, E., & Moller, S. (2018). The role of high school racial composition and opportunities to learn in students' STEM college participation. *Journal of Research in Science Teaching, 55*(3), 446–476. https://doi.org/10.1002/tea.21426

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Deconinck, E., Hancock, T., Coomans, D., Massart, D. L., & Heyden, Y. V. (2005). Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. *Journal of Pharmaceutical and Biomedical Analysis, 39*(1), 91–103. https://doi.org/10.1016/j.jpba.2005.03.008

Eccles, J. (1983). Female achievement patterns: Attributions, expectancies, values, and choice. *Journal of Social Issues, 1*, 1–26.

Eccles, J. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist, 44*(2), 78–89. https://doi.org/10.1080/00461520902832368

Feelders, A. (1999). Handling missing data in trees: Surrogate splits or statistical imputation? *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 329–334). Springer.

Franco, M. S., & Patel, N. H. (2017). Exploring student engagement in STEM education: An examination of STEM schools, STEM programs, and traditional schools. *Research in the Schools, 24*(1), 10–30.

Gottfried, M. A., & Bozick, R. (2016). Supporting the STEM pipeline: Linking applied STEM course-taking in high school to declaring a STEM major in college. *Education Finance and Policy, 11*(2), 177–202. https://doi.org/10.1162/EDFP_a_00185

Howard, N. R., Howard, K. E., Busse, R. T., & Hunt, C. (2019). Let's talk: An examination of parental involvement as a predictor of STEM achievement in math for high school girls. *Urban Education, 58*(4), 586–613. https://doi.org/10.1177/0042085919877933

Hudson, M. A., Baek, Y., Ching, Y. H., & Rice, K. (2020). Using a multifaceted robotics-based intervention to increase student interest in STEM subjects and careers. *Journal for STEM Education Research, 3*(3), 295–316. https://doi.org/10.1007/s41979-020-00032-0

Kitchen, J. A., Sonnert, G., & Sadler, P. M. (2018). The impact of college-and university-run high school summer programs on students' end of high school STEM career aspirations. *Science Education, 102*(3), 529–547. https://doi.org/10.1002/sce.21332

Kundu, M. G., & Harezlak, J. (2019). Regression trees for longitudinal data with baseline covariates. *Biostatistics & Epidemiology, 3*(1), 1–22. https://doi.org/10.1080/24709360.2018.1557797

Lee, A. (2015). Determining the effects of computer science education at the secondary level on STEM major choices in postsecondary institutions in the United States. *Computers & Education, 88*, 241–255. https://doi.org/10.1016/j.compedu.2015.04.019

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337–346. https://doi.org/10.1002/sim.3782
Lee, S. W., Min, S., & Mamerow, G. P. (2015). Pygmalion in the classroom and the home: Expectation's role in the pipeline to STEMM. *Teachers College Record, 117*(9), 1–40. https://doi.org/10.1177/016146811511700907
Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior, 45*(1), 79–122. https://doi.org/10.1006/jvbe.1994.1027
Matusovich, H. M., Streveler, R. A., & Miller, R. L. (2010). Why do students choose engineering? A qualitative, longitudinal investigation of students' motivational values. *Journal of Engineering Education, 99*(4), 289–303. https://doi.org/10.1002/j.2168-9830.2010.tb01064.x
Mau, W. C. J., & Li, J. (2018). Factors influencing STEM career aspirations of underrepresented high school students. *The Career Development Quarterly, 66*(3), 246–258. https://doi.org/10.1002/cdq.12146
Means, B., Wang, H., Young, V., Peters, V. L., & Lynch, S. J. (2016). STEM-focused high schools as a strategy for enhancing readiness for postsecondary STEM programs. *Journal of Research in Science Teaching, 53*(5), 709–736. https://doi.org/10.1002/tea.21313
Moakler, M. W., Jr., & Kim, M. M. (2014). College major choice in STEM: Revisiting confidence and demographic factors. *The Career Development Quarterly, 62*(2), 128–142. https://doi.org/10.1002/j.2161-0045.2014.00075.x
National Center for Education Statistics (2018). High school longitudinal study of 2009 (HSLS:09): User manuals. Retrieved December 15, 2020, from https://nces.ed.gov/surveys/hsls09/usermanuals.asp
National Science Foundation (2019). *Science and engineering indicators: post–high school transitions.* Retrieved January 24, 2021, from https://ncses.nsf.gov/pubs/nsb20196/post-high-school-transitions
National Science Foundation (2020). *STEM education for the future: a visioning report.* Retrieved June 15, 2022, from https://www.nsf.gov/ehr/Materials/STEM%20Education%20for%20the%20Future%20-%202020%20Visioning%20Report.pdf
Park, J. H., Lee, I. H., & Cooc, N. (2019). The role of school-level mechanisms: How principal support, professional learning communities, collective responsibility, and group-level teacher expectations affect student achievement. *Educational Administration Quarterly, 55*(5), 742–780. https://doi.org/10.1177/0013161X18821355
Pike, G. R., & Robbins, K. (2019). Expanding the pipeline: The effect of participating in project Lead the way on majoring in a STEM discipline. *Journal for STEM Education Research, 2*(1), 14–34. https://doi.org/10.1007/s41979-019-00013-y
Riegle-Crumb, C., King, B., Grodsky, E., & Muller, C. (2012). The more things change, the more they stay the same? Prior achievement fails to explain gender inequality in entry into STEM college majors over time. *American Educational Research Journal*, *49*(6), 1048–1073. 10.3102%2F0002831211435229
Riegle-Crumb, C., & Morton, K. (2017). Gendered expectations: Examining how peers shape female students' intent to pursue STEM fields. *Frontiers in Psychology, 8*, 329. https://doi.org/10.3389/fpsyg.2017.00329
Robinson, K. A., Perez, T., Carmel, J. H., & Linnenbrink-Garcia, L. (2019). Science identity development trajectories in a gateway college chemistry course: Predictors and relations to achievement and STEM pursuit. *Contemporary Educational Psychology, 56*, 180–192. https://doi.org/10.1016/j.cedpsych.2019.01.004
Saw, G. K., & Agger, C. A. (2021). STEM pathways of rural and small-town students: Opportunities to learn, aspirations, preparation, and college enrollment. *Educational Researcher, 50*(9), 595–606. https://doi.org/10.3102/0013189X211027528
Saw, G., & Chang, C. N. (2018). Cross-lagged models of mathematics achievement and motivational factors among Hispanic and non-Hispanic high school students. *Hispanic Journal of Behavioral Sciences, 40*(2), 240–256. https://doi.org/10.1177/0739986318766511
Saw, G., Chang, C. N., & Chan, H. Y. (2018). Cross-sectional and longitudinal disparities in STEM career aspirations at the intersection of gender, race/ethnicity, and socioeconomic status. *Educational Researcher, 47*(8), 525–531. https://doi.org/10.3102/0013189X18787818
Steinberg, D., & Colla, P. (2009a). CART: classification and regression trees. *The top ten algorithms in data mining*. X. Wu & V. Kumar (Ed.) 179–201. New York: CRC Press.
Steinberg, D., & Colla, P. (2009b). CART: Classification and regression trees. In X. Wu & V. Kumar (Eds.), *The top ten algorithms in data mining* (pp. 179–201). CRC Press.

Stets, J. E., Brenner, P. S., Burke, P. J., & Serpe, R. T. (2017). The science identity and entering a science occupation. *Social Science Research, 64*, 1–14. https://doi.org/10.1016/j.ssresearch.2016.10.016

Therneau, T. M., & Atkinson, E. J. (1997). *An introduction to recursive partitioning using the rpart routine*. (Vol. 61, p. 452) Mayo Foundation: Technical report.

Timofeev, R. (2004). *Classification and regression trees (CART) theory and applications (Unpublished doctoral dissertation)*. Humboldt University.

U.S. Bureau of Labor Statistics. (2017). *STEM occupations: past, present, and future*. Retrieved Jun. 1, 2021, from https://www.bls.gov/spotlight/2017/science-technology-engineering-and-mathematics-stem-occupations-past-present-and-future/home.htm

U.S. Census Bureau. (2019). *2019 American community survey, 1-year estimates*. Retrieved Jun. 23, 2022, https://www.census.gov/library/stories/2021/06/does-majoring-in-stem-lead-to-stem-job-after-graduation.html

U.S. Department of Education. (2015). *Science, technology, engineering and math: education for global leadership.* Retrieved Jun. 1, 2021, https://www.ed.gov/sites/default/files/stem-overview.pdf

U.S. Department of Education. (2016). *STEM 2026: a vision for innovation in STEM education.* Retrieved Jun. 1, 2021, from https://www.air.org/system/files/downloads/report/STEM-2026-Vision-for-Innovation-September-2016.pdf

Verbyla, D. L. (1987). Classification trees: A new discrimination tool. *Canadian Journal of Forest Research, 17*(9), 1150–1152. https://doi.org/10.1139/x87-177

Wang, X. (2013). Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal, 50*(5), 1081–1121. https://doi.org/10.3102/0002831213488622

Wille, E., Stoll, G., Gfrörer, T., Cambria, J., Nagengast, B., & Trautwein, U. (2020). It takes two: expectancy-value constructs and vocational interests jointly predict STEM major choices. *Contemporary Educational Psychology, 61*, 101858. https://doi.org/10.1016/j.cedpsych.2020.101858

Zheng, J. (2004). Study on the relationship of training data size to error rate and the performance comparison for two decision tree algorithms. [Doctoral dissertation, Texas Tech University].