From the
AERA Online Paper Repository
http://www.aera.net/repository

Paper Title How Rule Induction Data Mining Can and Cannot Be Useful for Education Research

Author(s) Emi Iwatani, Digital Promise

Session Title Data Mining and Unconventional Data Sources

Session Type Roundtable Presentation

Presentation Date 4/9/2019

Presentation Location Toronto, Canada

Descriptors Data Analysis, Data use, Statistics

Methodology Conceptual/Theoretical

Unit Division D - Measurement and Research Methodology

DOI 10.302/1438441

Each presenter retains copyright on the full-text paper. Repository users should follow legal and ethical practices in their use of repository material; permission to reuse material must be sought from the presenter, who owns copyright. Users should be aware of the AERA Code of Ethics.

Citation of a paper in the repository should take the following form: [Authors.] ([Year, Date of Presentation]). [Paper Title.] Paper presented at the [Year] annual meeting of the American Educational Research Association. Retrieved [Retrieval Date], from the AERA Online Paper Repository.

1969 words (excluding title & abstract)

How Rule Induction Data Mining Can and Cannot be Useful to Education Research

Abstract

This paper shares insights and recommendations on how rule induction data mining can and cannot be useful to education research, based on re-analyzing two regression studies with rule induction approaches. Processes and findings were compared to identify whether, how, and why rule induction could add value. I found that rule-based approaches can provide unique descriptions of the sample that shows at-a-glance, how key predictors relate to each other and to the outcome. They can also identify relationships between variables that held for some subgroups but not others. It was important to clearly understand the difference between *mining rules* and *mining rulesets*, as well as the unique research questions that these answer, so that they complement rather than replace regression.

Objective

Can rule induction data mining be useful in educational research? Rule induction methods—including decision trees, association rule mining and sequential covering—are a very popular subset of data mining approaches because of their flexibility, robustness to outliers, and ease of use and understanding. Most distinctively, rule induction approaches can identify interesting relationships among predictors and outcomes that apply only to a subset of the data (Hand, 1997; Witten, Frank, & Hall, 2011) — relationships that may even contradict the general trend. This paper shares key findings from an experiential and theoretical exploration on how this underutilized quantitative approach may add methodological value to education research.

Perspective/theoretical framework

While some educational researchers have eschewed exploratory quantitative research as "data-fishing," I aligned myself with those who see value in it (Grover & Mehra, 2008; Zhao & Luan, 2006). At the same time, I attended to traditional concerns about threats to statistical and ecological validity of inferences, as innovations are unlikely to be adopted unless they align with the values of the field (Rogers, 2003).

Furthermore, drawing from the diffusion of innovations theory (Rogers, 2003), I assumed that a methodology wouldn't be useful unless it provides information that cannot be gained, or cannot easily be gained, by existing approaches. I also assumed that rule induction would be less likely to be adopted if it is perceived as too complex, or "relatively difficult to understand and use" (Rogers, 2003, p.257).

Methods

I re-analyzed two regression studies on the National Educational Longitudinal Study of the Eight Grade Class of 1988 (NELS:88) dataset using rule induction approaches, comparing results and process across methods. One used hierarchical regression to identify predictors of science achievement (Byrnes &

Miller, 2007), while the other used logistic regression to identify predictors of Black students' academic success of Black students (Thomas, 2006). I conducted rule induction algorithms first by including only the predictors in their final models, then by using all reasonable predictors in NELS:88. Ruleset induction approaches included: RIPPER (Cohen, 1995), CBA (Liu, Hsu, & Ma, 1998; Liu, Ma, & Wong, 2000), PART (Frank & Witten, 1998), CART (classification and regression tree, Breiman, Friedman, Olshen, & Stone, 1984), C4.5 (Quinlan, 1993) and See5/C5.0 (Quinlan, 2013) and QUEST (Quick Unbiased Efficient Statistical Tree, Loh & Shih, 1997). Apriori (Agrawal, Imieliński, & Swami, 1993) was also used, adjusted so that it only generates rules with the outcome variable as a consequent. Random forests, bagged version of CART and boosted version of C4.5 were included (although they do not result in rules) to further contextualize the predictive accuracies of the other algorithms. A hold-out approach was used to avoid model over-fit, creating the model with 70% of the data, and testing with the remaining. The 70-30 random split was stratified across outcome, and the same split was used across algorithms. R v.3.3.2 (R Core Team, 2016) and RStudio v.1.0.136 (RStudio Team, 2015) were used for most the data mining, although SAS software v.9.4 and SPSS v.22 for some cleaning and analyses. I mostly used a SONY VAIO laptop v.1511, with an Intel(R) CORE™i7 processor (2GHz), 8GB RAM and 64-bit operating system.

Rule and ruleset induction happened in stages (Figure 1). I first generated rulesets from the NELS data, then created representations of the output to make it easier to interpret, then finally gleaned findings and possible implications. For both studies, the ruleset findings helped decide subgroups for which to conduct association rule mining by suggesting variables (including cut-scores of numeric variables) that were strongly related to the outcome. I then conducted similar processes for association rule mining—going from data to output (steps 1 and 5), to representations of output (steps 2 and 6), then to insight (steps 3 and 7). Across analyses of each study, I compared the predictive accuracy (confusion matrices, F-ratio), relative importance of predictors included in the model, and examined potentially interesting predictor-outcome relationships that could be unique among subgroups. Finally, I reflected on the results and process to gain insight on rule induction's relative advantage and practicality – this reflection is the focus of this paper.

Data

Selected findings are displayed in Tables 1-7, and Figures 1-8, with key findings from each re-analysis summarized in Table 9 and Table 10. "Study 1" refers to the re-analysis of Thomas (2006), while "study 2" refers to that of Byrnes & Miller (2007). This paper's focus, however, is not the data outputs and summary *per se*, but rather the insights gained from reflecting on this analysis and meaning-making process.

Results

Rulesets vs rules

Ruleset induction is better suited for modeling rather than identifying specific rules. This is because as a newspaper article loses much of its meaning when viewed in isolation from the context in which it was created, rules lose the relational meaning against its context when viewed in isolation from the ruleset within which it was created. For example, a C4.5 rule associating 12^{th} grade math scores with high 8^{th} grade math achievement, geometry, and high math self-concept, becomes more meaningful when examining the entire model in which that rule is embedded (Figure 6). The context indicates the order in which the predictor variables were associated with the outcome, and how frequently specific attribute-values appear in other rules. For this example, the rule by itself suggests that interventions on 8th grade math, geometry and math self-concept may help a small group of individuals, while the ruleset additionally suggests that interventions on 8th grade math is helpful for everyone while interventions on math self-concept may matter to just a smaller portion of students, and so on.

Thus, the ruleset approach is less suited for identifying interesting individual rules because of its unexhaustive nature of the rule search, and built-in dependency among rules. In other words, rules within a ruleset are less reliable because they are not only a function of the predictors' relationship to the outcome, but also a function of the search algorithm and other rules within the ruleset. Association rule mining, on the other hand, conducts an exhaustive search for rules, and each rule it generates is independent of the search algorithm or of other rules that were identified within the ruleset. Thus, association rule mining would be more helpful than ruleset modeling for identifying interesting rules, and indeed it turned out as such with this project.

How rule induction does and does not add value beyond traditional statistical approaches

In both studies, rule induction models had comparable or slightly worse predictive accuracies relative to regression (Figure 2, Figure 3). This is unsurprising given the truism in machine learning that there is no universally superior algorithm (Wolpert, 2012; Wolpert & Macready, 1997). However, rulesets can provide researchers with a unique description of the sample, different from a regression-based picture, that shows at-a-glance, how some of the key predictors were related to the outcome and to each other. This descriptive value was particularly vivid for the models in Study 2, where the mosaic diagrams characterized the sample and their 12th grade math scores first in terms of 8th grade math scores, then mainly in terms of math course-taking (e.g., Figure 6, Figure 7). The difference in descriptive power is largely because rulesets describe every respondent, while regression tables describe average contributions by individual variables (an abstraction, so more difficult for the mind to imagine). Ruleset results were also clearer than regression in expressing how just a handful of variables were generally sufficient to explain most of the explainable outcome variance. Byrnes and Miller provided analogous information through presenting partial correlations of each predictor with the outcome, but in general, relative contributions of predictors to the outcome are somewhat difficult to ascertain in regression unless one knows how to interpret results tables. Thus, the model limitations are perhaps more visible with rulesets, which can keep researchers from overextending their inference.

Rules and rulesets also can help identify relationships between variables that held for some subgroups but not others. For example, rulesets in Study 2 suggested that Algebra 2 and math self-concept were positively related to 12th grade math scores, but only for those who were higher achieving in 8th grade math, while general math was negatively related to 12th grade math scores but only for those who scored lower on 8th grade math (Table 10). Association rule mining provided similar types of findings. In study 1, for example, several factors (e.g., participation in honors or gifted and talented programs, school safety) were more strongly associated with 12th grade achievement for lower income students, and students whose parents did not have a college degree (Table 9). In contrast, regression provides how each predictor contributes, on average, to the population.

In addition, rule induction could identify cut-points of continuous predictors, and groupings of nominal predictors, that could be useful for prediction and further analysis. For example, the 8th grade math subgroups identified by CART in study 2 (Figure 7) motivated the next step of using those scores as cut-points for creating subsamples to conduct association rule mining.

Rule and ruleset induction also easily identified variables that are related to the outcome that were not included in the regression model. However, inclusion of unexpected variables in the model is not a unique benefit of ruleset induction *per-se*, but rather a feature of any kind of data mining that makes its variable choice explicit. Stepwise regression, for example, likewise may have identified different variables.

Finally, a point that is seemingly obvious, but not stated clearly enough so far in the literature, is that rule induction approaches are not directly helpful in answering the research questions that regression is specifically designed to answer. Ruleset induction answers an exploratory and sequential question: *What is a set of characteristics that tend to be commonly associated with each level of the outcome, and to those to whom that set of characteristics do not apply, what (if any) is another set of characteristics that would apply, (and so on)?* Regression answers: *Assuming every independent variable has the same amount of impact on the individual after controlling for other factors, how much unique impact does each independent variable have?* So instead of whether and which variable significantly predicts the outcome, rule induction is better equipped to characterizing *to whom* a factor matters for predicting an outcome, and identifying factors that those with similar levels of outcome have in common. In assessing ruleset results, I was tempted to focus on the relative predictor importance, wondering about the average effect predictors had on the population (Figure 4 & Figure 5). But regression better answers those questions.

Recommendations

There appear to be at least three practical and principled ways to incorporate rule induction into education research in the future, in ways that complement rather than conflate it with regression. (1) Use of ruleset mining to describe the sample, and how some of the key independent and dependent variables relate. This was the way ruleset induction was used in this study, and leads to descriptions of how characteristics associated with different level of subgroups. (2) Use of association rule mining to identify what factors, if any, are different across groups. The groups could reflect differences in outcome (e.g., high achieving vs low achieving), in treatment, or background. This was the way association rule mining was used in this study. (3) Use of decision trees to identify whether a predictor or predictors are related to the outcome, *after controlling for key covariates*. Like hierarchical regression, the idea would be to first model the outcome with a set key of covariates, then at each of the terminal nodes to investigate the relationship(s) between the outcome and the independent variables(s) of most interest using e.g., regression or a non-parametric approach. There are algorithms that instantiate some versions of this, such as the logistic model tree that conducts logistic regression at each terminal node of a decision tree.

Tables and figures

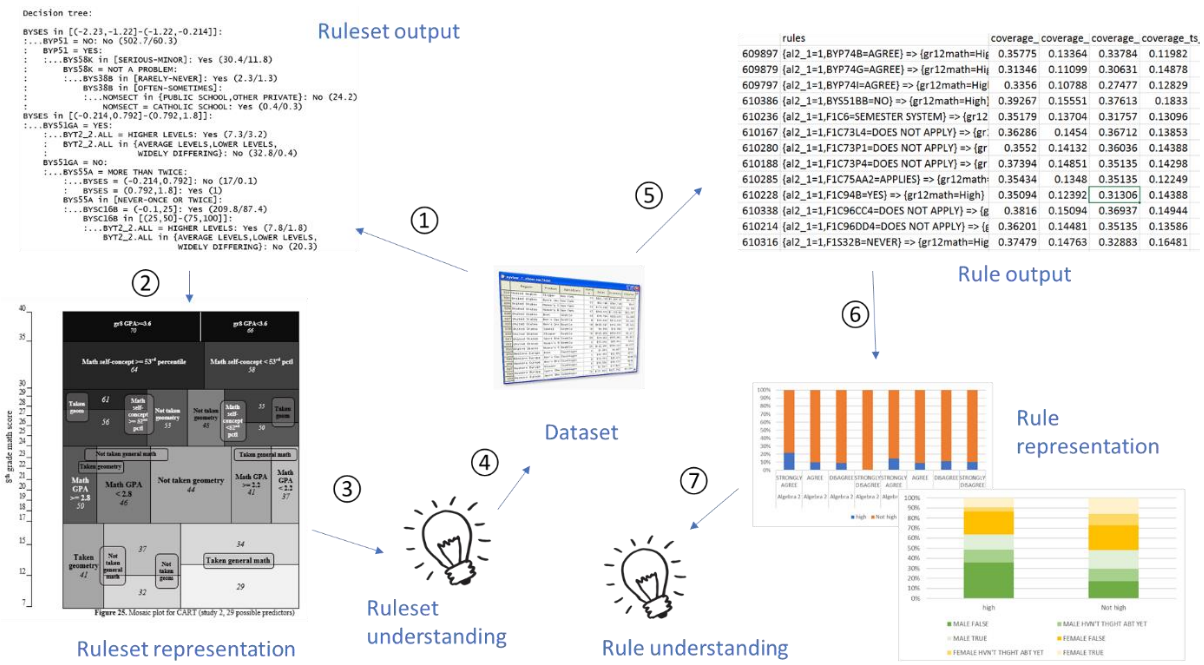

Figure 1. Illustration of rule and ruleset mining process for project

Table 1. Logistic regression prediction of Black student achievement with Thomas' (2006) final variables

Variable	Replication		Thomas (2006)	
	B	*Exp(B)*	*B*	*Exp(B)*
Intercept	-2.622***	.073	-1.344***	.261
Parental education (BYPARED)	.415***	1.514	.201**	1.223
Income from all sources 1991 (F2P74)	.067*	1.069	.083**	1.087
Hours of homework in school (F2S25F1)	.044	1.045	.039	1.040
Hours of homework out of school (F2S25F2)	.128**	1.137	.255***	1.290
Household resources (hhressc)	.029	1.029	.038	1.039
Private school (privsch)	-.668^{+}	.512	-1.4**	.247
Religious school (religsch)	.677	1.969	1.775**	5.900
Parental involvement in school (pinvolve)	.050	1.051	.069	1.071
Parents expect college (pexpcol)	.081	1.084	.82***	2.270
Good peers (goodpeer)	-.078	.925	-.549**	.578
Bad peers (badpeer)	-.313^{+}	.731	-.774***	.461
Peers expect college (peerexcl)	.342*	1.408	.357*	1.429
Activities outside of school (activity)	.016	1.016	.055	1.057
Student's cultural activities (sculture)	.131*	1.140	.086	1.090
Percent receiving free lunch in school (G8LUNCH)	-.097*	.907	-.051	.950
School climate (climate)	.026	1.026	.07**	1.073
Student feels unsafe in school (unsafe)	-.799**	.450	-.801**	.449
Disruptions in school prevent learning (disrupt)	-.573***	.564	-.421*	.656
Number of Black, non-Hispanic teachers (BYSC20D)	-.114**	.892	-.116**	.890

Table 2. Hierarchical regression on 12th grade math achievement with Byrnes & Miller's (2007) variables

Variable	Replication Coef	Replication ΔR^2	Byrnes & Miller Coef	Byrnes & Miller ΔR^2
Distal factors		.428		.430
SES in 8th gr (BYSES)	.821***		.790***	
Parent expectations in 8th gr (Pexp)	.445***		.478***	
Student expectations in 8th gr (Sexp)	.445**		.409**	
Middle school GPA (BYGRADS)	1.182***		1.215***	
Opportunity factors		.112		.112
General math ½ yr (gm_half)	-2.401**		-2.826**	
General math 1yr (gm_1)	-3.052***		-3.338***	
General math 1.5-2yrs (gm_2)	-3.078***		-3.477***	
Geometry ½ yr (geo_half)	1.452***		1.620***	
Geometry 1 yr (geo_1)	2.544***		2.660***	
Geometry 1.5-2yrs (geo_2)	2.156		2.237	
Algebra II ½ yr (al2_half)	1.591**		1.374**	
Algebra II 1yr (al2_1)	1.207***		.858**	
Algebra II 1.5-2yrs (al2_2)	.446		.261	
S perception of math emphasis (emph_m)	.111		.157*	
S perception of T responsiveness (t_rspnsv)	.232*		.132*	
Propensity factors		.224		.219
Math achiev before 9th gr (BYTXMIRR)	.975***		.678***	
Math GPA in 9th & 10th gr (GPA910_m)	.978***		.943***	
Efficacy for graduating HS (grad_eff)	.704**		.915**	
Plans to take SAT (SATplan)	1.161***		1.180***	
Math self-concept (m_selfcpt)	.310***		.175***	
Demographic factors		<.01		<.01
Female	-1.822***		-1.725***	
Black	-1.845***		-2.184***	
Hispanic	-.594		-.761*	
Asian	.877		.779	
Native American	-1.740		-2.078	

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; N=8976 for replication, while 8969 for Byrnes & Miller. ΔR^2 result from a hierarchical regression where the distal factors are entered, followed by opportunity factors, propensity factors and demographic factors. The coefficient estimates are of the model that includes all predictors. The opportunity factors alone accounted for 45.1% (45.2% per Byrnes and Miller) of the outcome variance, propensity factors accounted for 73.1% (72.4%), and gender and race/ethnicity factors 9.1% (9.3%).

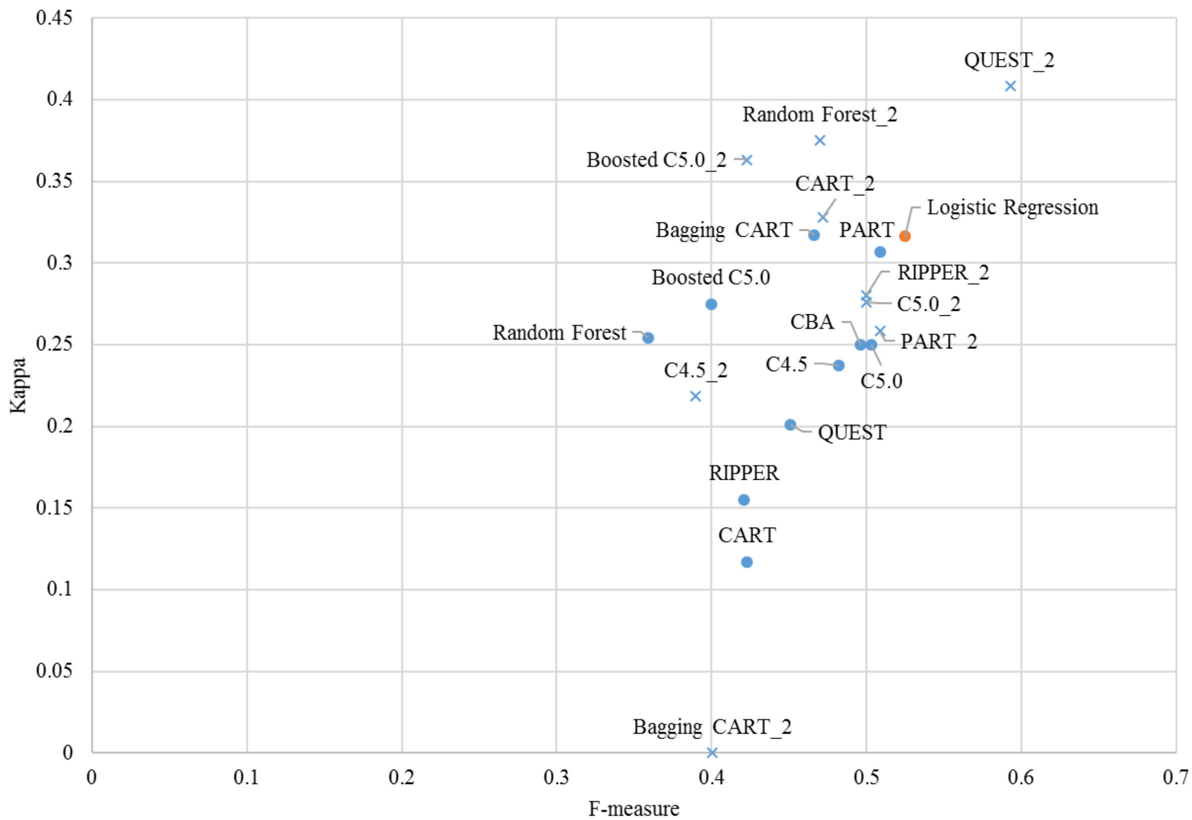

Figure 2. F-measure and Kappa statistics of logistic regression vs rule induction (Study 1)[1]

[1]The orange marker represents result from logistic regression, while the blue markers indicate results from rule induction approaches. Circle markers indicate results using Thomas' (2006) 19 possible predictors, while "X"s indicate results using 1372 possible predictors.

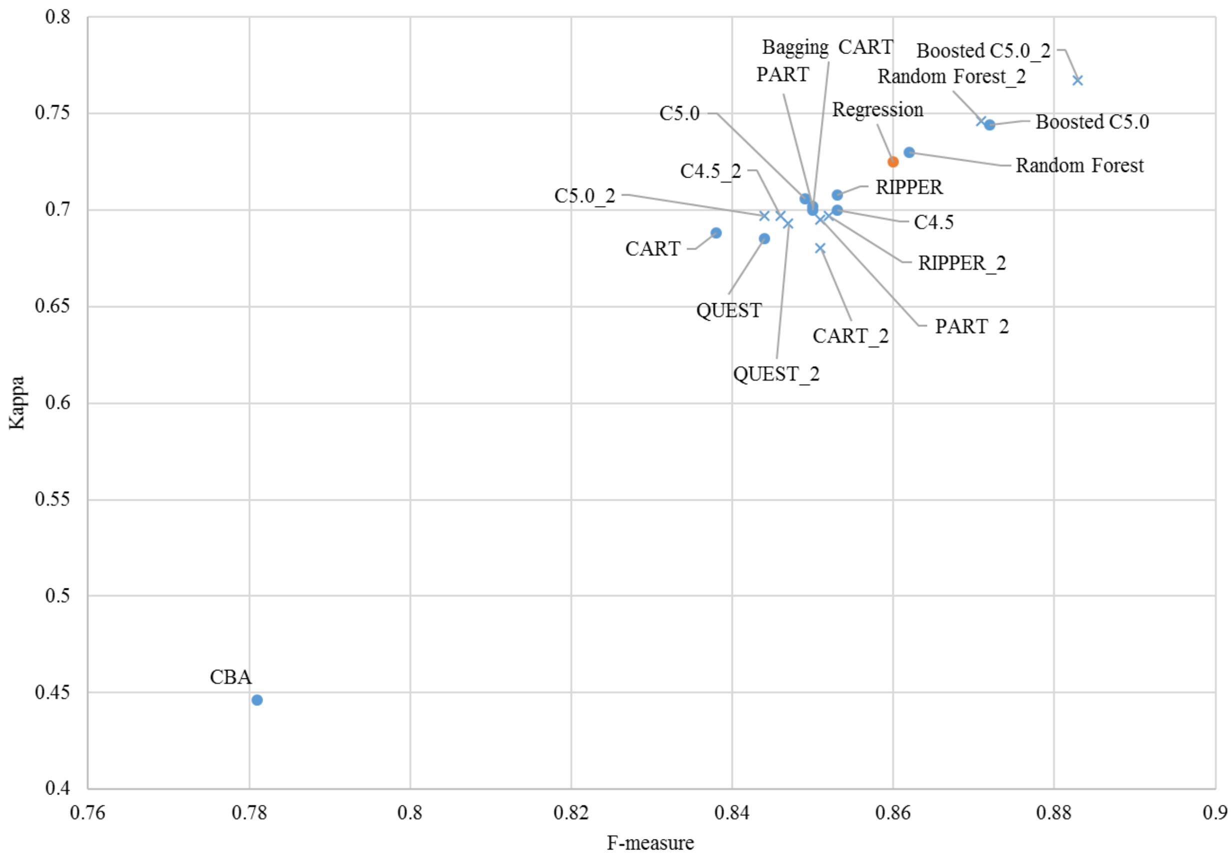

Figure 3. F-measure and Kappa statistics of logistic regression vs rule induction (Study 2)[1]

[1]The orange marker represents result from regression, while the blue markers indicate results from rule induction approaches. Circle markers indicate results using Byrnes & Miller's (2007) 29 possible predictors, while "X"s indicate results using 1933 possible predictors.

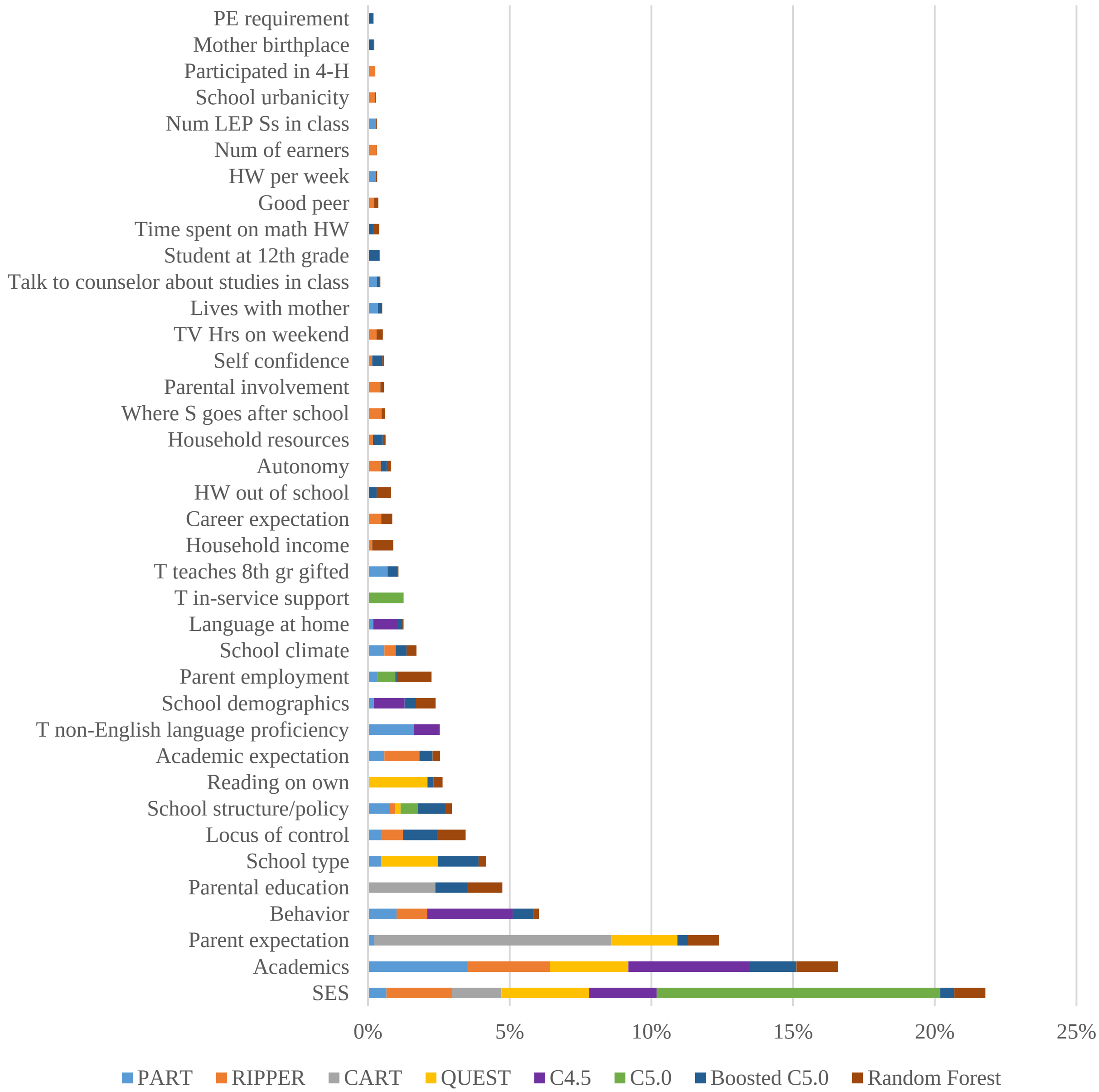

Figure 4. Predictor importance by algorithm (Study 1, 1372 possible predictors)

Note: Predictor importance for CART and Random Forest were calculated as the extent to which the variable reduced the Gini index, as these were automatically generated. For the rest of the algorithms where that measure was not available, attribute usage was used. Attribute usage was calculated as the number of participants that the variable sorted, where participants were double-counted as being sorted by that variable if, in tree-induction, the variable was used twice to sort the same person and if there was at least one other variable that sorted that person between the first and second sorting. The predictor importance for each algorithm was scaled to total 100, predictors that contributed less than 1 in all of the non-ensemble algorithms were excluded from consideration, and remaining predictors were grouped into the categories in the figure.

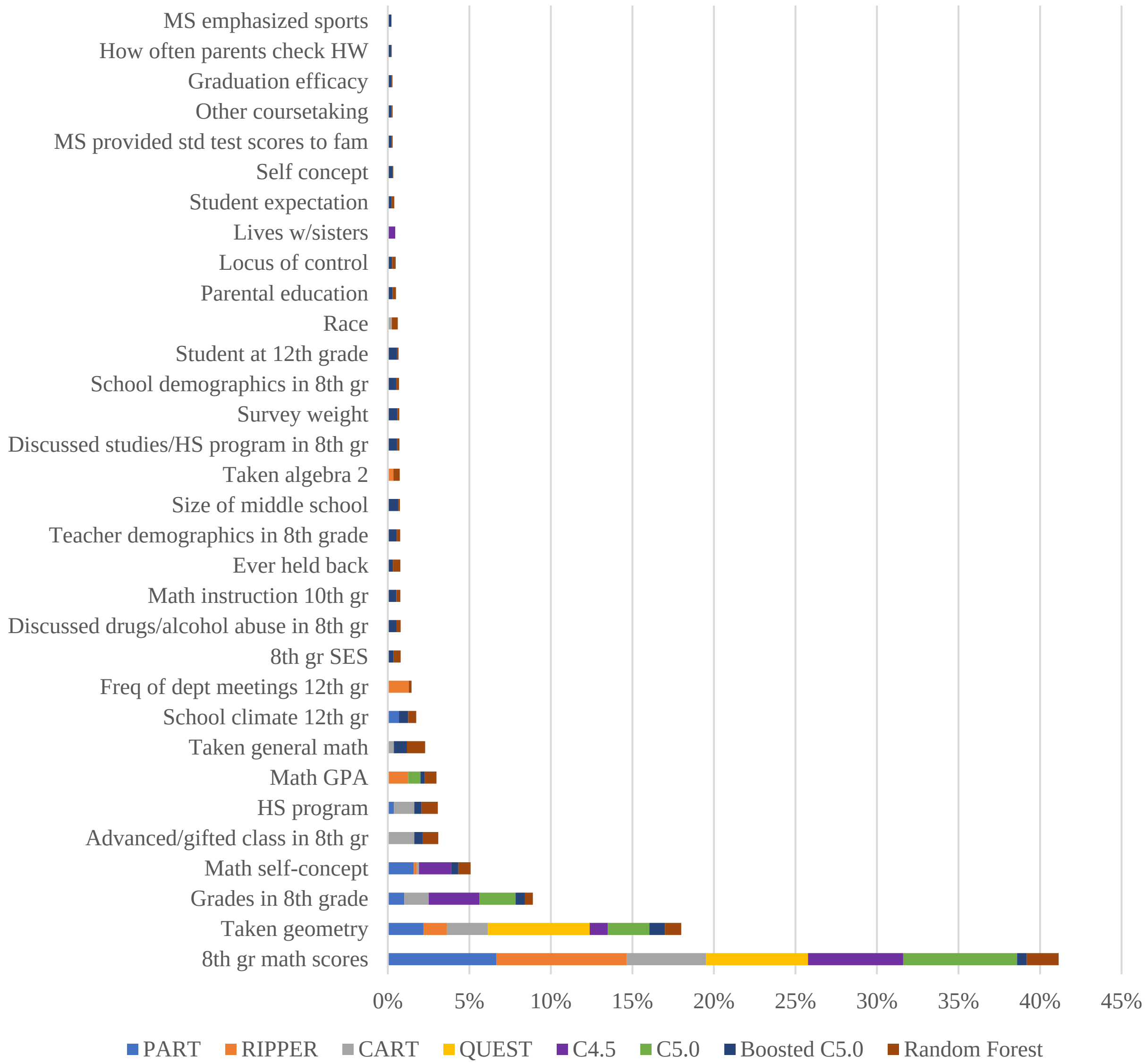

Figure 5. Predictor importance by algorithm (Study 2, 1933 possible predictors)

Note: Predictor importance for CART and Random Forest were calculated as the extent to which the variable reduced the Gini index or mean squared error, respectively, as these were automatically generated. For the rest of the algorithms where that measure was not available, attribute usage was used. Attribute usage was calculated as the number of participants that the variable sorted, where participants were double-counted as being sorted by that variable if, in tree-induction, the variable was used twice to sort the same person and if there was at least one other variable that sorted that person between the first and second sorting. The predictor importance for each algorithm was scaled to total 100, predictors that contributed less than 1 in all of the non-ensemble algorithms were excluded from consideration, and remaining predictors were grouped into the categories in the figure.

Table 3. Number of rules, interesting rules, and false alarms discovered by algorithm (Study 1)

	Variables considered	Rules	Interesting	False alarm
CBA	19	67	*	*
RIPPER	19	7	1	
	1372	18	3	1
PART	19	31		
	1372	16	2	
C4.5	19	8		1
	1372	23	1	1
CART	19	2		
	1372	2		
C5.0	19	12	1	
	1372	14	2	2
QUEST	19	5		
	1372	8	1	
Total		213	11	5

* Note: Interestingness analysis was not conducted for CBA due to intractable number of rules.

Table 4. Number of rules discovered (Study 2)

	Variables considered	Rules
CBA	19	832
RIPPER	19	6
	1372	6
PART	19	6
	1372	10
C4.5	19	7
	1372	12
CART	19	20
	1372	15
C5.0	19	9
	1372	9
QUEST	19	10
	1372	4
Total		946

Table 5. Interesting rules discovered by ruleset induction (Study 1)

Set of predictors & rule origin	**Consequent**	**Reasons for interest**	**Reason for disinterest or caution in interpretation**
Been held back in ES or MS (C4.5, large, R21 & R22)	Low achieving	Predictor not considered by Thomas. Applies quite widely (13%/15%)* and is fairly accurate in classifying those whom the rule applies (95%/93%, where random chance would predict 75%). First predictor to be selected by algorithm, indicating strength of prediction of low achievers relative to others. Same predictor was selected across multiple algorithms. Potentially significant implications for educational theory and practice.	Might already be well-known by the field.
Peers expect college, no disruptions and parental involvement is 68th percentile or above (RIPPER, small, R1)	High achieving	Specific combination of predictors not highlighted by Thomas. Applies quite widely (11%/12%)* and is fairly accurate in classifying those whom the rule applies (50%/56%, where random chance would predict 25%). First rule to be selected by algorithm, indicating strength of prediction relative to others. Potentially significant implications for educational theory and practice.	Might already be well-known by the field.

Set of predictors & rule origin	Consequent	Reasons for interest	Reason for disinterest or caution in interpretation
SES is less than 89th percentile, not taking higher level courses in 8th grade, and parents are not doctors, and (but) read on own 6hr+/wk in 8th grade. (QUEST, large, R2)	High achieving	Reading and course-taking predictors were not considered by Thomas. Predictive validity and generality are reasonable (see next column). Potentially significant implications for educational theory and practice. Follow-up analysis found that the voracious reader condition applied to 5%/6% of those whose SES is less than 89th percentile, not taking higher level courses in 8th grade, and parents are not doctors, and when applied, was correct 36% (vs 14%/13% chance).	Rule does not apply very widely (3%), and is not extremely accurate (42%/33%, where random chance would predict 25%).
SES is higher (64th-93rd percentile), has been counseled about drug/alcohol and sent to the office more than twice in 8th grade. (C5.0, large, R9)	Not high achieving	Behavioral factors were not considered by Thomas. High predictive accuracy (100%)	Applies to a small group of students (2%/0.8%) Might be well-known in field.
If SES is higher (64th+ percentile), and behavior is good (haven't spoken to counselor about drug/alcohol abuse and have not been sent to office more than twice), and parent had thought about S's test scores being probably not good enough to qualify for loan/scholarship (i.e., selected "true" or "false" rather than "haven't thought about it") (C5.0, large, R11)	High achieving	Behavioral factors and 8th graders' parents' beliefs about college financial aid were not considered by Thomas. Applies quite widely (21%/22%)* and is fairly accurate in classifying those whom the rule applies (59%/50%, where random chance would predict 25%).	[Follow-up analysis] The contributions of the behavior and financial aid variables to the prediction are somewhat small (coverage was 71% and lift was 1.32/1.18 when examining only those whose SES is 64th+ percentile)

Set of predictors & rule origin	**Consequent**	**Reasons for interest**	**Reason for disinterest or caution in interpretation**
Parental educational attainment is more than high school and less than a 4-year college degree, parent expects college, student does not feel unsafe in school and fewer than 10 Black teachers in school. (C5.0, small, R8)	High achieving	Specific combination of predictors not highlighted by Thomas. Applies quite widely (22%/21%)* and somewhat accurate in classifying those to whom the rule applies (34%/35%, where random chance would predict 25%). [Follow-up analysis] The rule was even more predictive when examining only those whose parental educational attainment was more than high school and less than college degree, with coverage of 33%/32% and confidence of 34%/35% where random would predict 22%. Potentially significant implications for educational theory and practice.	Predictive validity is not extremely high. May not have practically significant implications for research and practice.
[Excluding 12% of respondents who were predicted to be high achieving by the first rule based on a combination of SES, academic, family and school demographic factors] If student was not held back, does not have a math teacher who teaches gifted & talented program, does not attend a religious school, and locus of control is at or lower than 37th percentile. (PART, large, R2)	Low achieving	Locus of control and whether math teacher teaches gifted/talented program were not considered by Thomas. Applies quite widely (28%/24%)* and somewhat accurate in classifying those to whom the rule applies (93%/87%, where random chance would predict 71%/78%). Potentially significant implications for educational theory and practice.	Predictive accuracy is not as compelling in the test set (lift – 1.10, as opposed to 1.30 in training set). Might be well-known in field. May not have practically significant implications for research and practice.

Set of predictors & rule origin	Consequent	Reasons for interest	Reason for disinterest or caution in interpretation
[Excluding 23% of respondents who were predicted to be low achieving by the first 4 rules based mainly on sense of safety and parent education] Doing 7-12 hours of homework outside of school, and peers expect college (PART, large, R5 & R6)	High achieving	Specific combination of predictors not identified by Thomas as being more predictive than others [Follow-up analysis] Applies reasonably widely (17%/14% for just the homework condition, 14%/9% for both homework and peer college expectation conditions)* and somewhat accurate in classifying those to whom the rule applies (51%/45% for just the homework condition where random chance would predict 35%/32%, and 55%/53% for both conditions where random chance would predict 30%/28%). Potentially significant implications for educational theory and practice.	Reduced generality and accuracy in test set. Might be well-known in field. May not have practically significant implications for research and practice.
SES is 58th percentile or higher, in higher achieving classes and never sent to office for misbehavior (RIPPER, large, R1)	High achieving	Includes predictors not considered by Thomas. Specific combination of predictors not identified by Thomas as being more predictive than others Applies somewhat widely (7%) and quite accurate in classifying those to whom the rule applies (87%/67%, where random chance would predict 25%).	Reduced generality and accuracy in test set. Might be well-known in field. May not have practically significant implications for research and practice.

Set of predictors & rule origin	Consequent	Reasons for interest	Reason for disinterest or caution in interpretation
[Not in above high achieving group] SES is 36th percentile or higher, parent has expectations for PSE and parent does not think child's test scores will be too low to qualify for college financial aid. (RIPPER, large, R2)	High achieving	Includes predictors not considered by Thomas. Applies somewhat widely (10%/12%) and quite accurate in classifying those to whom the rule applies (71%/49%, where random chance would predict 25%). Potentially significant implications for educational theory and practice.	Reduced generality and accuracy in test set. Might be well-known in field. May not have practically significant implications for research and practice.
[Not in above two high achieving groups] Locus of control is 54th percentile or above, autonomy is 56th percentile or higher, watches over 5 hours of TV a day in the weekend, student goes home after school, and household has one income earner. (RIPPER, large, R3)	High achieving	Includes predictors not considered by Thomas. Applies somewhat widely (2%/4%) and quite accurate in classifying those to whom the rule applies (79%/47%, where random chance would predict 25%).	Reduced generality and accuracy in test set. Might be well-known in field. May not have practically significant implications for research and practice.

*"Small" and "large" refers to dataset with 19 and 1372 predictors, respectively. "R##" refers to the rule number indicated in Appendix C.
**Percentages in parentheses indicate predictive validity for the training set (first percentage) and test set (second percentage). Only one number (the more conservative of the two) is indicated if the two were within a percentage point difference.

Table 6. Factors associated with high achievement among parental education and income subgroups (Study 1)

Category	**Details**	**Groups to which rules applied** (PLR=relative probability)
[Student] Enrolled in gifted/talented program in 8th grade	Enrolled in gifted/talented program in 8th grade according to parent or student. Participated in academic honors society.	**Low parental education** (TPR .32-.26; PLR=2.2-3.5/2.1-4.8) **Low income** (TPR .25-.37/.29-.25; PLR=2.3-3.5/2.9-5.6) Note: Parent report more predictive than student report for gifted/talented program participation.
[Student] Attends higher level classes in 8th grade	Taking algebra, in higher ability group for math and/or English. Does not attend regular math.	**Low parental education** (TPR=.41-.53/.48-.68; PLR=1.9-3.0/1.7-2.8) **High parental education** (TPR=.49-.53/.28-.47; PLR=1.8-2.3/1.5-1.7) **Low income** (TPR .41-.52; PLR=2.0-3.0/2.9-4.0) **High income** (TPR .43-.57/.32-.51; PLR=2.2-3.5/1.5-4.5) Note: For high parental education and high-income groups, math only (not English) was predictive. Prediction for high parental education group was not as strong relative to other groups.
[Student] T believes Ss class is higher achieving than average	T considers achievement of Ss class to be higher achieving relative to average, according to English, Science and average of all 4 core subject area teachers.	**Low parental education** (TPR=.21-.4/.12-.36; PLR=3.3-5.5/1.7-2.8) **Low income** (TPR .25-.42/.35-.48; PLR=3.9-6.3/2.6-6.7)
[Student] High locus of control in 8th grade	Highest quartile on 2 types of locus of control composite. Disagrees that chance/luck is important in life.	**Low parental education** (TPR=.44-.62/.43-.64; PLR=1.9-2.7 /1.9-2.1) **Low income** (TPR=.29-.52/.32-.64; PLR=1.6-3.0 / 1.7-2.3) **High income** (TPR=.47-.57/.45-.68; PLR=1.7-2.4 / 1.7-1.8)
[Student] 8th grader expects postsecondary education after high school	Expects to go to school after high school. Expects to attend a college prep program after high school.	**Low parental education** (TPR=.39-.49/.38-.45; PLR=2.0-2.4/2.1-2.3) **Low income** (TPR=.44-.47/.42-.45; PLR=2.1-2.5/1.9-2.5) **High income** (TPR=.46-.47/.55-.61; PLR=2.4-3.1/2.2-2.6)
[Student/Family] Parents expect 8th grader to get more schooling after college	Mother expects higher school after college. Father expects higher school after college.	**High income** (TPR=.38-.50/.48-.52; PLR=.17-.16/.19)

[Student/Family] Parent expects child's HS test score to not be bad	Parent does *not* believe that their child's test score would not be good enough for child to qualify for financial aid	**Low parental education** (TPR = .70/.78; PLR=1.7/1.6) **Low income** (TPR = .78/.74; PLR=1.8/1.5) **High income** (TPR = .78/.68; PLR=1.7/1.5)
[Student] 8^{th} grader studies music	Attends music at least once a week. Participated in band or orchestra. Child studies music outside regular school.	**Low parental education** (TPR=.53/.57; PLR=1.6/2.1) **Low income** (TPR=.27-52; PLR=1.7-2.5/1.5-2.0)
[Student] 8^{th} grader studies foreign language	Enrollment/attendance in foreign language course.	**Low parental education** (TPR=.25-.27/.19-.21; 2.0-2.5/1.7-1.8) **High income** (TPR=.40/.49; PLR=2.6/2.4)
[Student/family] 8^{th} grader does not work for pay	8^{th} grader does not work for pay	**High income** (TPR = .39/.42; PLR=1.8/2.5)
[Student] 8^{th} grader expects a professional, managerial or business occupation	8^{th} grader expects a professional, managerial or business occupation at age 30	**Low income (**TPR = .40/.42; PLR=1.9/1.6) **High income** (TPR = .40/.42; PLR = 1.7/1.7)
[School] 8^{th} grader's school has moderate attendance issues and minor other behavioral issues	Student tardiness and absenteeism are considered a "moderate" problem (in a scale of "serious", "moderate", "minor" and "not a problem") by student. Robbery/theft, and verbal abuse of teachers at school are considered either a "minor" and/or "moderate" problem by student. Class cutting is considered "not a problem" by teacher or student.	**Low parental education** (TPR=.20-.47/.26-.36; PLR=1.6-2.1/1.5-2.2) **Low income** (TPR=.26-.37/.19-.42; PLR=1.6-4.1/1.4-3.3) Note: Class cutting rule applied to high income group as well.
[School] 8^{th} grader's math class emphasizes Algebra	Algebra is a major topic in student's math class.	**Low income** (TPR=.34/42; PLR=1.5)
[School] 8^{th} grader's school has formal admissions procedures	School has formal admissions procedures.	**Low parent education** (TPR=.23/.24; PLR=2.0) **High income** (TPR= .38/.39; PLR=1.9/2.3)
[School] 8^{th} grader is challenged at school	Parent "strongly agrees" that their 8^{th} grader is challenged at school.	**Low parent education** (TPR=.32/.31; PLR=2.1/2.6) **Low income** (TPR= .32/.39; PLR=2.1/3.6)
[School] Parent believes 8^{th} grader's school sets realistic standards	Parent "strongly agrees" that school sets realistic standards.	**Low parent education** (TPR=.27/.29; PLR=1.6/2.0)
[School] 8^{th} grader feels safe at school	8^{th} grader "strongly disagrees" that they do not feel safe at their school	**Low parent education** (TPR=.44/.52; PLR=1.5/1.8)

		Low income (TPR= .49/.55; PLR=1.7/1.8)
[School] Parent is very satisfied with 8th grader's education	Parent "very satisfied" with education child has received	**High parent education** (TPR=.68/.59; PLR=2.1/1.6)
[Family] Parent of 8th grader goes to museums	Parent reports going to history, art, and/or science museums	**Low parent education** (TPR=.26-.27/.33-.34; PLR=1.6/1.7-1.9) **Low income** (TPR=.26-.32/.23-.32; PLR=1.6-2.4/1.7-3.7)
[Student] 8th grader goes to history museums	Parent reports that 8th grader goes to history museums	**Low parent education** (TPR=.43/.47; PLR=1.6/1.9) **Low income** (TPR=.47/.35; PLR=2.7/1.7)

TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving.

Table 7. Association between 12th grade achievement and participation in 8th grade gifted/honors program by income and parental education subgroups (Study 1)

Group	Var name	Proportion of gifted students among high achievers (TPR)	Proportion of gifted students among lower achievers (FPR)	TPR / FPR (PLR)	Proportion of high achievers among gifted (Precision)	Proportion of high achievers among not-gifted (FOR)	Precision / FOR (RP)
Low income	BYS68A	.38	.15	2.53	.35	.13	2.62
	BYS82O	.26	.08	3.25	.4	.14	2.81
	BYP51	.33	.08	4.13	.46	.13	3.51
Low parental educ	BYS68A	.32	.15	2.13	.2	.08	2.34
	BYP51	.26	.07	3.71	.3	.08	3.57

TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving. For high parental education group, TPR .61-.62; PLR=.73-.87. For high income group, TPR .57-.63; PLR=.66-.91.

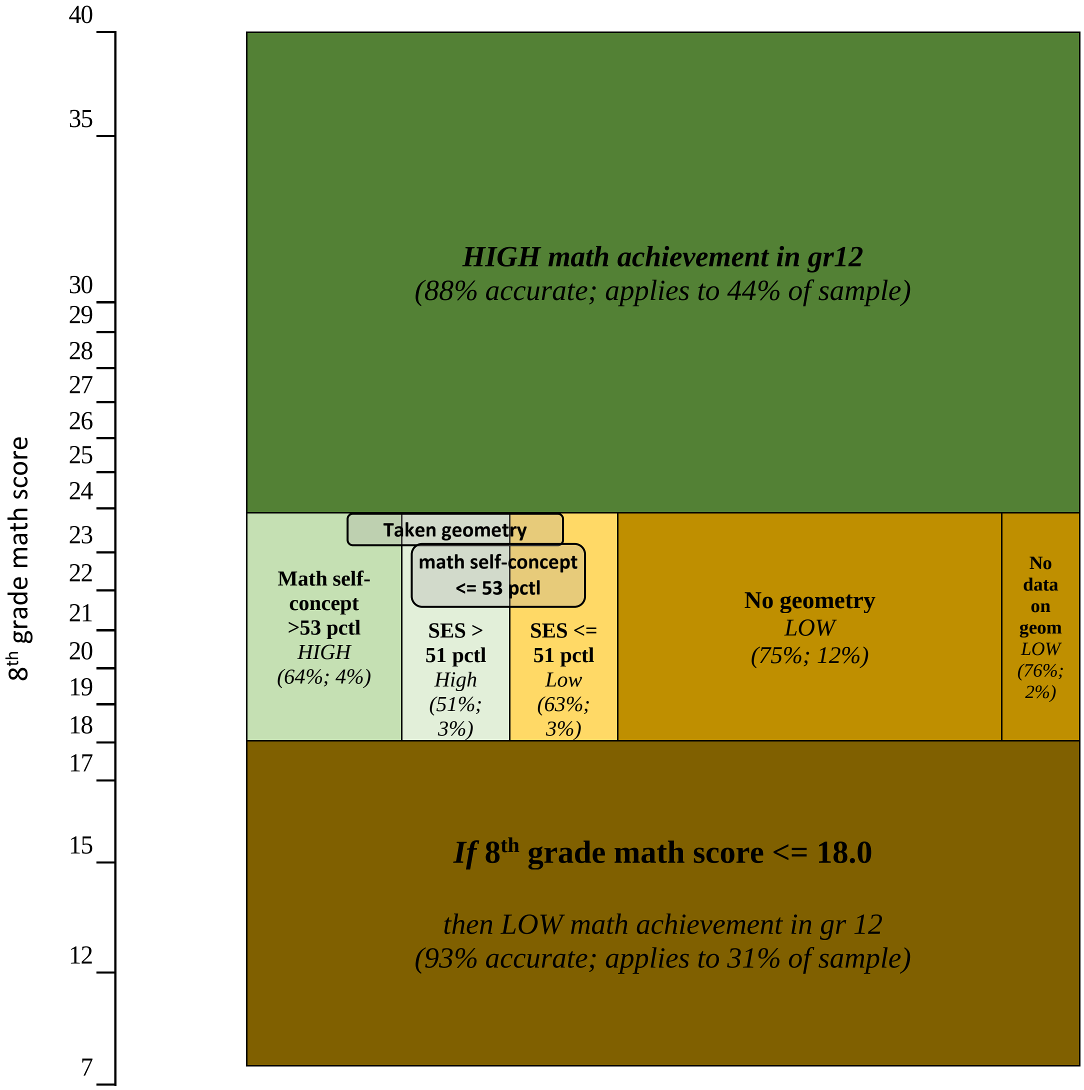

Figure 6. Mosaic plot for C4.5 (Study 2, 29 possible predictors)

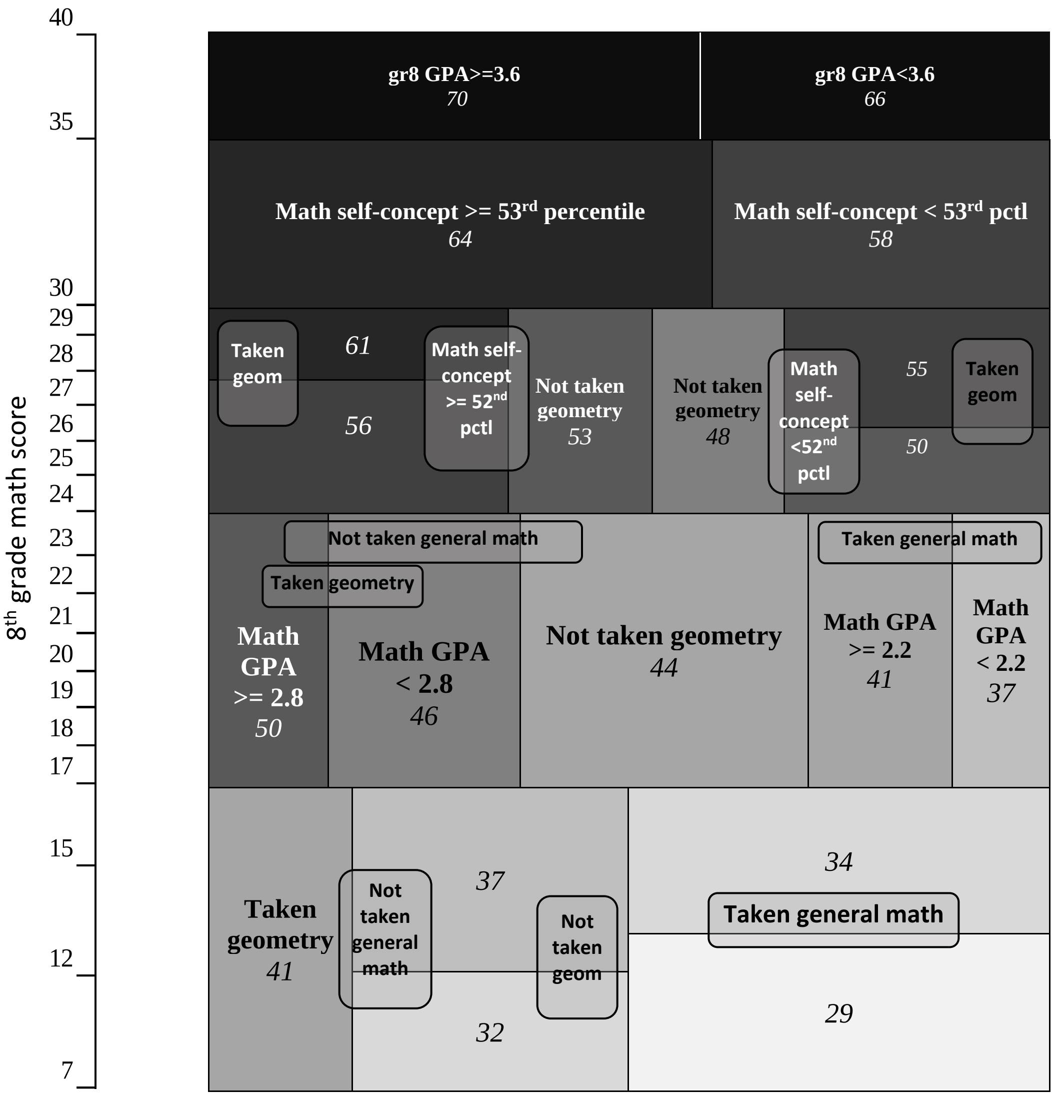

Figure 7. Mosaic plot for CART (Study 2, 29 possible predictors)

Table 8. Variables associated with higher than expected math achievement in 12th grade, within 3 different 8th grade math achievement subgroups (Study 2)

Category	**Details** (square brackets refer to subgroups for which the rule was discovered, where 1 through 3 are groups with 8th grade math scores <17, 17-23 and 24-31, respectively)	**Groups to which rules applied** (TPR = true positive rate, PLR=positive likelihood ratio)
[Opportunity] Math course-taking in HS	Taken geometry. [1] Taken algebra 2. [2, 3]	**Gr8 math score <17** (TPR = 0.31/0.38; PLR = 1.8/2.3) **Gr8 math score 17-23** (TPR = 0.27/0.38; PLR = 1.7/2.3) **Gr8 math score 24-31** (TPR = .43-.49/.44-.5; PLR = 2.1-1.9/1.8-1.9)
[Opportunity] HS emphasis in academics	S attends academic HS program; "Very accurate that HS Ss are expected to do HW; 75-100% of HS Ss in academic HS program. [1] S attends academic HS program; T "agrees" that dept is committed to AP and honors programs. [2] Ss write science labs once a week [3]	**Gr8 math score <17** (TPR = .29-.44/.28-.43; PLR = 1.5-1.8/1.6-1.9) **Gr8 math score 17-23** (TPR = .31-.50/.40-.61; PLR = 1.6-1.7/1.7-1.9) **Gr8 math score 24-31** (TPR = 0.48/0.52; PLR = 1.5/1.5)
[Opportunity] School safety and climate	HS teacher considers robbery/theft, illegal drugs, alcohol and possession of weapons to be a "minor" problem at the school.	**Gr8 math score 17-23** (TPR = .32-.43/.35-.47; PLR = 1.6-1.9/1.5-2)
[Opportunity/other] Teachers have necessary materials	HS teacher "agrees" that necessary materials are readily available.	**Gr8 math score <17** (TPR = 0.29/0.43; PLR = 1.5/2) **Gr8 math score 17-23** (TPR = 0.46/0.53; PLR = 1.6/1.7)
[Opportunity/other] Teachers in a positive, learning-oriented culture	HS teacher "agrees" that they are encouraged to experiment with teaching, grading practices are consistent and fair, and/or department chair consults staff before decision. Teacher "disagrees" that routine practices interfere with teaching. Teacher reports that cooperative learning and higher-order thinking skills are discussed.	**Gr8 math score 17-23** (TPR = .31-.51/.25-/53; PLR = 1.6-1.9/1.5-1.8)
[Distal] Parent expects college	Parent expects 8th grader to attend a 4-5 year college program.	**Gr8 math score <17** (TPR = 0.42/0.38; PLR = 1.7/1.5)
[Propensity] P believes 8th grader's academics will not negatively interfere with college financial aid	Parent does not expect 8th grader's test scores and/or grades to be too low to qualify for financial aid.	**Gr8 math score <17** (TPR = .47-.51/.46-.44; PLR = 1.6-1.7/1.6-1.7)
[Propensity] P does not believe that they have not	8th grader's parent does not believe they have not been able to get much	**Gr8 math score <17** (TPR = 0.35/0.31; PLR = 1.6/1.6)

been able to get information about how to apply for financial aid.	information on how and where to apply for financial aid.	
[Distal] Parent expectation for algebra	8th grader believes their parents/guardian wanted 8th grader to take Algebra.	**Gr8 math score <17** (TPR = 0.3/0.4; PLR = 1.6/2)
[Opportunity] Enrollment in foreign language class	8th grader enrolled in a foreign language class.	**Gr8 math score 24-31** (TPR = 0.39/0.51; PLR = 1.5/1.7)
[Other] Availability of cheerleading	Cheerleading available to 8th graders at the school.	**Gr8 math score <17** (TPR = 0.35/0.32; PLR = 1.6/1.5)
[Other] 8th grader has been threatened once or twice	"Once or twice," someone has threatened to hurt 8th grader at school.	**Gr8 math score 24-31** (TPR = 0.29/0.28; PLR = 1.6/1.5)
[Distal] Parent/guardian was in mid 30s when 8th grader was born	Parent who responded to base year survey was born in 1940-1944 (48-52 years old in 1988; i.e., 34-38 years old when 8th grader was born).	**Gr8 math score 24-31** (TPR = 0.26/0.26; PLR = 1.6/1.5)
[Opportunity/other] Social studies teacher's teaching was observed several times by supervisor.	Social studies teacher reports that supervisor observed their teaching "several times."	**Gr8 math score <17** (TPR = 0.28/0.27; PLR = 1.7/1.5)

Table 9. Key findings from Study 1

Method	Findings
Logistic regression	Factors positively associated with 12th grade achievement included: Higher parental education, more homework out of school, higher family income (females), lower percentage of students receiving free/reduced lunch, fewer disruptions in school, greater sense of safety in school, fewer Black teachers. And also in Thomas's analysis: Parental expectation for college, having "good" peers, positive school climate.
Ruleset mining with 19 predictors	Parental education sufficient to determine much of the explainable variance. After parental education, school safety, number of Black teachers in school, and household income tend to explain much of the remaining explainable variance.
Ruleset mining with 1372 predictors	Higher SES, higher 8th grade academic achievement, higher parent expectation, and fewer behavioral issues in 8th grade account for much of the explainable variance in the outcome. Parental education, school type, locus of control, school structure/policy, reading on own and students' academic expectations explained much of the remaining explainable variance.
Examination of rules within rulesets	16 of 213 rules appeared to provide additional information to regression in the training set, but 5 seemed to be false alarms based on test set. Combinations of predictors predicted outcome well for some students. Factors not selected by Thomas were included in the expanded model. Was practically not feasible to calculate validity metrics for CBA rules.
Association rules—rules for household income (high, low) and parental education (high, low) subgroups	Participation in honors or gifted and talented programs, attending higher-than-average level classes, parents' strong agreement that student is challenged at school, school safety, and study of music were more strongly associated (PLR >= 2.5) with higher achievement in 12th grade for lower income students and students whose parents did not have a college degree. For those whose parents had a college degree and those from higher income households, participation in honors/gifted programs was more strongly associated with 12th grade achievement (PLR >=4) when in conjunction with some other factor. For the higher parental education group, there were many such factors, which tended to pertain to attendance in stable/positive schools and programs, or having good behavior. For those from higher income households, there were only three such factors, which seemed conceptually unrelated to one another (e.g., discipline is fair, friends do not impact students' decision to take algebra). If the student had taken a higher-level math class in 8th grade and/or not taken a remedial math, their likelihood of being in the high achieving group was 40-50%, which was 1.7 to 3.2 times the likelihood of the remaining sample being high achieving. The positive likelihood ratio was increased to at least 4 if another factor—such as school safety, high locus of control, and good behavior, parental expectation that students' test scores would be good enough for college financial aid—also applied. Parents' assessment of the school seemed to be a greater factor for students from lower income families. Students' perception of their educational environment (of teachers, policies, usefulness of schoolwork) and SES seemed to be a greater factor for students from higher income families. Direct question to parent about expectation for college-after-high-school was associated with 12th grade achievement only for students from higher household income, while indirect question to parent and direct question to student were associated with outcome regardless of income. Not working for pay associated with higher achievement only for students from higher income families.

Table 10. Key findings from Study 2

Method	Key findings
Hierarchical multiple regression	Model accounted for 76% of variance (with over 90% of that sufficiently explained by math self-concept; and with most variables accounting for <1% of unique variance in the outcome). Distal factors (incl., SES, parent & student expectations and middle school GPA) explained 43% of 12th grade math score variance, opportunity factors (esp., math course-taking) explained 11% more (or 45% by itself), propensity factors (esp. MS math achievement) explained 22% more (or 73% by itself). Demographic factor explained less than 1% of remaining variance. Following factors had high correlations with outcome variable: each of the four distal factors such as 8th grade SES and parent expectations (correlation between .42 and .56), 1 year of general math (-.37), 1 year of geometry (.53), and 1 year of Algebra II (.37), math achievement before the start of 8th grade (.84), GPA in grade 9 and 10 (.44), and math self-concept (.40)
Ruleset mining with 29/1933 possible predictors	Predictive accuracy was comparable to regression. 8th grade math scores most predictive of outcome, followed by math course-taking. 8th grade math scores are the most important predictor across all ruleset models, included in every rule. When outcome is dichotomized, additional factors (primarily math course-taking and sometimes other factors such as middle school GPA, math self-concept, SES) improve the prediction, but generally not for those who score highest or lowest in 8th grade math. The CART mosaic plots—with the numeric outcome—indicated that 8th grade math scores were most predictive, and that what was next most predictive depended on that score. For those who scored the lowest in 8th grade math, geometry and general math course-taking were next most predictive. For those who scored higher, geometry course-taking and math self-concept were next most predictive. For those who scored even higher in 8th grade math, math self-concept was next predictive. For those who scored the highest, grade 8 GPA was next predictive, if anything.

Method	Key findings
Association rule mining	Math course-taking and HS emphasis in academics were associated with higher-than-predicted* achievement regardless of 8th grade scores, but there were slight differences on how these predictors were operationalized across different subgroups of 8th grade math scores. Re: Rules of length 2, with PLR >1.5: Parent's academic expectations for the 8th grader to take algebra, attend college, and/or qualify for college financial aid were associated with math achievement only for students whose 8th grade test scores were very low. School climate, safety, and teacher perception of their professional environment were associated with achievement only for students whose 8th grade scores were between 17 and 23. Some rules appeared less reliable (more one-off) than others. Re: Rules of length 3, with PLR >2.5: For students whose 8th grade math scores 24-31, algebra course-taking was included in every rule antecedent, some type of school factor (e.g., safety, worthwhileness of homework, availability of student clubs, and use of textbooks and hall passes) was included in many, while demographics, and peer or community variables were not included in any. Very few rules characterized those whose 8th grade math scores were lower than 24. For students with 8th grade math scores 17-23, the condition that the high school teacher agreed they are encouraged to experiment with teaching was true for 50% of the high achievers and 29% of others. If in addition, the 8th grader had visited with science/history museum or it was "very accurate" that the high school encourages students to take academic classes, the rule became true for about 30% of high achievers and just 12-15 % of lower achievers. For those with 8th grade math scores <17, being male and either attending a school that places emphasis on academics, or whose parents do not believe the students' test scores will be too low to qualify for college financial aid, were likely to be higher achieving than CART predicts. *Higher than predicted by CART, which primarily relied on 8th grade math scores.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). *Mining association rules between sets of items in large databases.* Paper presented at the ACM SIGMOD Record.

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, Calif: Wadsworth International Group.

Byrnes, J. P., & Miller, D. C. (2007). The relative importance of predictors of math and science achievement: An opportunity–propensity analysis. *Contemporary Educational Psychology, 32*(4), 599-629. doi:http://dx.doi.org/10.1016/j.cedpsych.2006.09.002

Cohen, W. W. (1995). *Fast effective rule induction.* Paper presented at the Proceedings of the twelfth international conference on machine learning.

Frank, E., & Witten, I. H. (1998). *Generating Accurate Rule Sets Without Global Optimization.* Paper presented at the Proceedings of the Fifteenth International Conference on Machine Learning.

Grover, L. K., & Mehra, R. (2008). The Lure of Statistics in Data Mining. *Journal of Statistics Education, 16*(1).

Hand, D. J. (1997). *Construction and assessment of classification rules*. Chichester; New York: Wiley.

Liu, B., Hsu, W., & Ma, Y. (1998). *Integrating classification and association rule mining.* Paper presented at the The 4th International Conference on Knowledge Discovery and Data Mining (KDD-98), Menlo Park, CA.

Liu, B., Ma, Y., & Wong, C.-K. (2000). *Improving an exhaustive search based rule learner.* Paper presented at the The 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000), Lyon, France.

Loh, W., & Shih, Y. (1997). Split selection methods for classification trees. *Statistica sinica, 7*(4), 815-840.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Mateo, Calif: Morgan Kaufmann Publishers.

Quinlan, J. R. (2013). See5/C5.0 (Version 2.10): Rulequest Research. Retrieved from http://www.rulequest.com/see5-info.html

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rogers, E. M. (2003). *Diffusion of innovations* (Vol. 5th). New York: Free Press.

RStudio Team. (2015). RStudio: integrated development for R.

Thomas, L. L. (2006). *Pathways to success or failure: Factors affecting academic achievement among black students.* (3203974 Ph.D.), State University of New York at Buffalo, Ann Arbor. Retrieved from http://pitt.idm.oclc.org/login?url=http://search.proquest.com/docview/304938982?accountid=14709 ProQuest Dissertations & Theses Full Text; ProQuest Dissertations & Theses Global database.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining : Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann.

Wolpert, D. H. (2012). *What the no free lunch theorems really mean: How to improve search algorithms*. Retrieved from http://www.santafe.edu/media/workingpapers/12-10-017.pdf

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67-82.

Zhao, C.-M., & Luan, J. (2006). Data Mining: Going beyond Traditional Statistics. *New Directions for Institutional Research*(131), 7-16.