CSIRO

Australia's National
Science Agency

Generation STEM: education data analytics project summary

Background

The Generation STEM initiative is funded by the New South Wales Government and delivered by CSIRO. The initiative aims to attract, support, and retain NSW students and works in STEM education and careers. As part of the initiative, a collaboration was established between CSIRO and the Catholic Education Parramatta Diocese (the Diocese). The collaboration was intended to better understand what factors are related to students pursuing STEM electives in high school and academic performance in STEM subjects. The project received human research ethics approval, and the Diocese provided a de-identified dataset to CSIRO to analyse. The dataset comprised subject, attendance, grades, school (de-identified) and some basic demographic information about students and their parents/carers for students in years 8 to 10, and HSC exam results for students in Year 12. There were several limitations with the dataset, including some missing data fields, many students entering and leaving the school system, and a limited number of calendar years of data (2017 to 2021). Although these issues reduced the robustness of the analyses (including two predictive analytics methods and a Bayesian network analysis), they still provided many useful insights.

Key findings

Predictive analytics using all factors

Predictive analytics were conducted using the subset of students with Year 12 HSC data (n = 2,007). The two target variables were:

- students who took high proportions of STEM subjects (50 per cent or more) in Year 12
- average grade in Year 12 STEM subjects

The target variables are the things that model is trying to predict.[1] Several machine learning models were run to determine the most

Table 1. The factors that best predicated becoming a Year 12 STEM student (highest feature values (SHAP))

1. NAPLAN Year 9 spelling score
2. NAPLAN Year 9 reading score
3. NAPLAN Year 9 writing score
4. Human Society and Environment Year 9 grade
5. Science Year 9 grade
6. Commerce Year 9 grade
7. Languages Year 9 grade
8. English Year 8 grade
9. NAPLAN Year 9 numeracy score
10. Commerce Year 10 grade

Table 2. The factors that best predicted grades in Year 12 STEM subjects (highest feature values (SHAP))

1. Science Year 10 grade
2. Personal Development, Health, and Physical Education Year 10 grade
3. Number of Year 12 STEM subjects taken
4. Human Society and Environment Year 10 grade
5. NAPLAN Year 9 numeracy
6. Mathematics Year 10 grade
7. Science Year 9 grade
8. Religious Education Year 10 grade
9. English (Advanced) Year 11-12 grade
10. English Year 10 grade

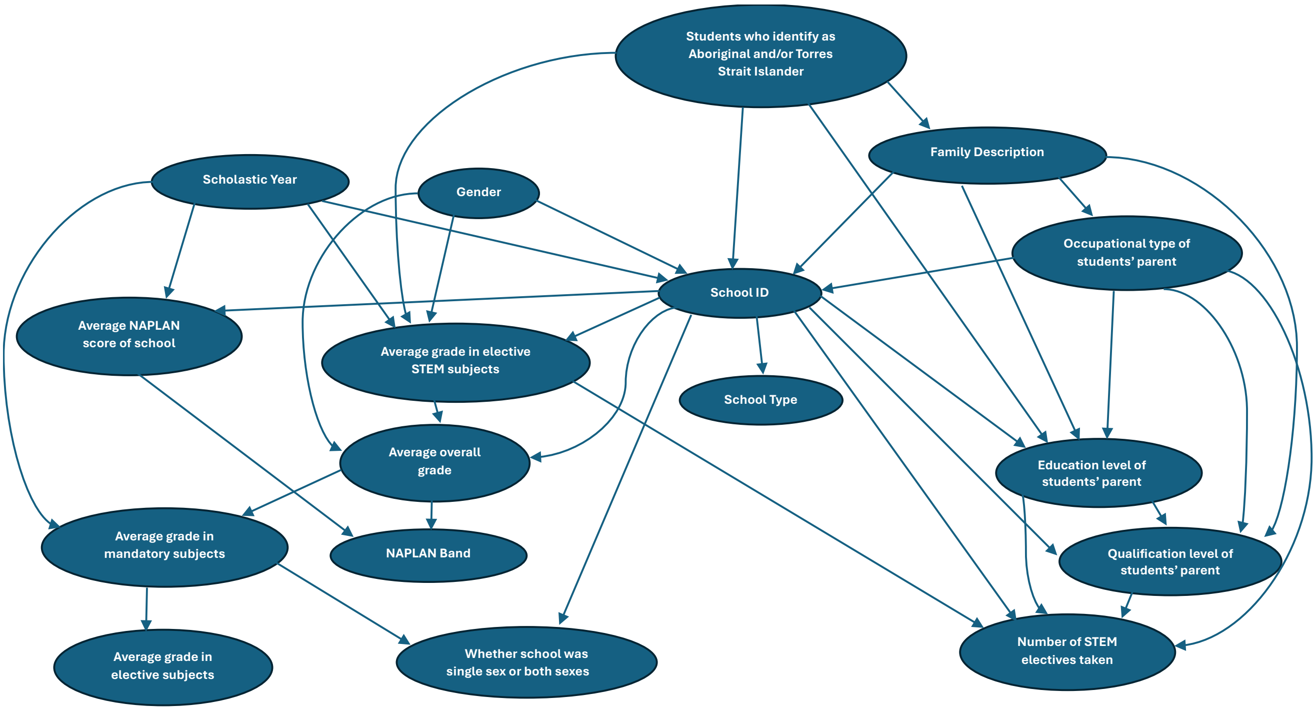
[1] In predictive analytics, independent variables, also called predictor, treatment, explanatory or feature variables, are the variables used to make inferences or predict the dependent variables, also called response, explained, or target variables.

influential features[2] for the prediction of the target variables. The features are the factors or variables for each student used to predict the target variables. The most important features for taking HSC STEM subjects (See Table 1) were Year 9 NAPLAN scores (Spelling, Reading, Writing), grades in Year 9 Science, grades in several non-STEM subjects in Years 8 through 10 (including HSIE, Commerce, Language, English), and Year 9 NAPLAN (Numeracy) scores. Being male and parents' education level were of relatively lower influence, which was somewhat unexpected. More in line with previous research, grades in Year 10 science and mathematics and the number of STEM subjects taken were of the most relative importance predicting grades in Year 12 STEM subjects (see Table 2). The most important features for predicting obtaining the highest grades in Year 12 STEM subjects were scores in Year 9 NAPLAN (Numeracy) and grades in Year 10 Science.

Bayesian Network analysis

A Bayesian Network model was developed using data from 5,119 students in Years 8 to 10. The model is a representation of variables and their relationships and includes some prior knowledge, for example what variables are likely or unlikely to be causally related. The dataset was streamlined by combining a number of variables. The first network (Figure 1) represents the relationships among a number of academic, demographic, and school variables. One variable of interest is 'Average grade in elective STEM subjects'; the model shows that the variables influencing average grades in STEM subjects are gender, Aboriginal and/or Torres Strait Islander status, and school attended. There are a number of other mediating relationships through the school, including parental education and qualification levels, and occupational group. This means that these other factors depend on what school the student attended.

Figure 1. Bayesian Network model 1 (Years 8 to 10 data)

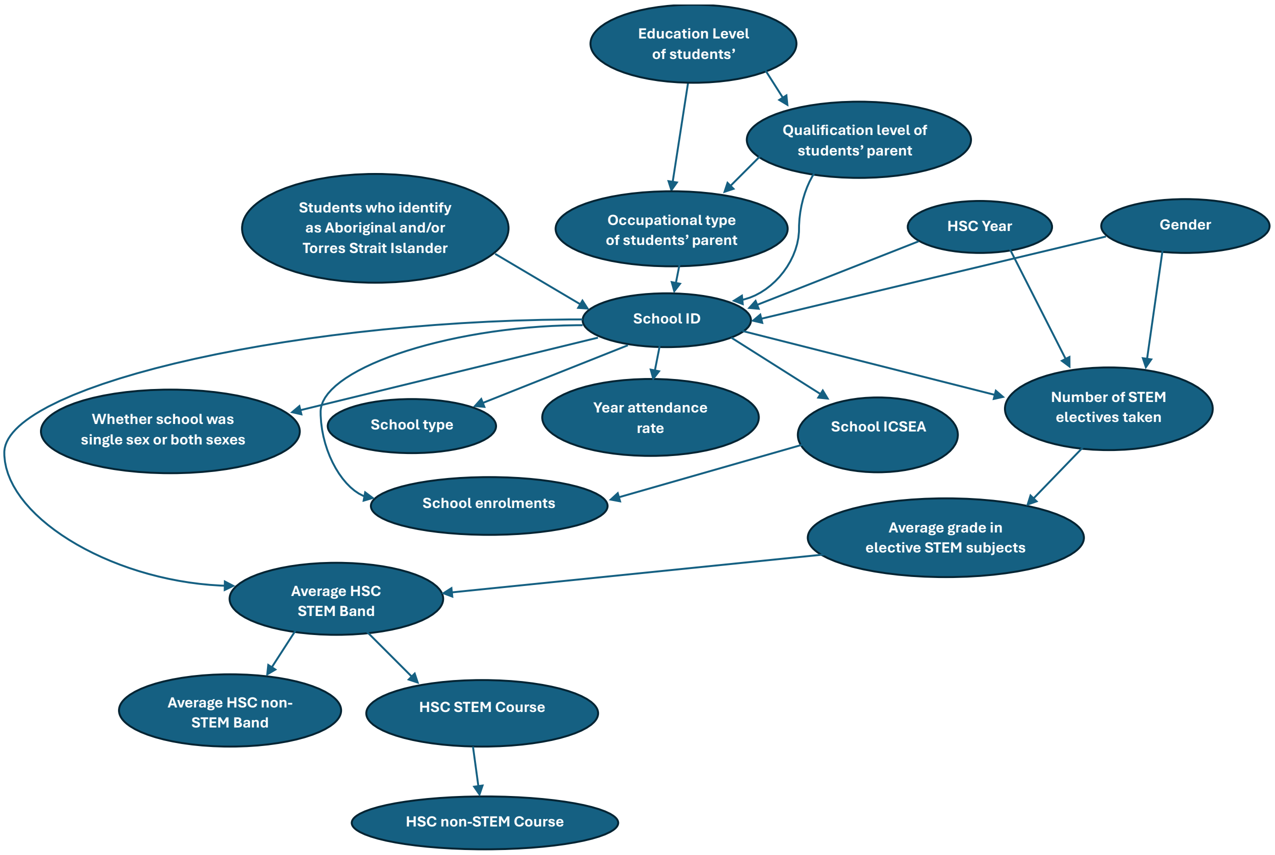

Students who identify as Aboriginal and/or Torres Strait Islander
Family Description
Scholastic Year
Gender
Occupational type of students' parent
School ID
Average NAPLAN score of school
Average grade in elective STEM subjects
School Type
Average overall grade
Education level of students' parent
Average grade in mandatory subjects
NAPLAN Band
Qualification level of students' parent
Average grade in elective subjects
Whether school was single sex or both sexes
Number of STEM electives taken

An additional model (Figure 2) focused on HSC Year 12 results using the data from 2,007 students. The model indicated that the key variables influencing grades in STEM subjects were grades in elective STEM subjects (in Year

[2] Feature importance was determined by SHAP (SHapley Additive exPlanations) values, which are a way to explain the output of a machine learning model, where each feature is assigned an importance value representing its contribution to the model's output.

10) and the school the student attended. Interestingly, there was not a substantial difference in average grades in STEM subjects among male and female students, and female students performed slightly better in non-STEM subjects.

Figure 2. Bayesian Network model 2 (HSC grades)


Education Level of students'
Qualification level of students' parent
Students who identify as Aboriginal and/or Torres Strait Islander
Occupational type of students' parent
HSC Year
Gender
School ID
Whether school was single sex or both sexes
School type
Year attendance rate
School ICSEA
Number of STEM electives taken
School enrolments
Average grade in elective STEM subjects
Average HSC STEM Band
Average HSC non-STEM Band
HSC STEM Course
HSC non-STEM Course

Predictive analytics for Year 10 results

Focusing on only Year 9 and 10 students in the dataset allowed for a larger sample of complete data (2,334 students with Year 9 and 10 data). Using Year 9 data (treatment variables) to predict Year 10 STEM grades (target variable) resulted in only a modestly accurate model.[3] However, it is important to note that the most important features (predictor variables) were in line with existing research, including: attendance, being male, the qualification and education levels of students' parents, and the ICSEA[4] value of the school (see Table 3). Considering attendance rate as the treatment variable, the average treatment effect[5] was significant, that is, for one standard deviation increase in attendance, there was a

Table 3. What best predicted average Year 10 STEM grades (highest feature values (SHAP))

1. Attendance rate
2. Gender (male)
3. Parent qualification
4. Parent education level
5. Parent occupation group
6. School ICSEA value
7. School size (enrolments)

[3] Specifically, the Root Mean Squared Error or RMSE (a performance indicator for regression models) of 0.77 and an R^2 (the coefficient of determination, a measure indicating the proportion of variance in the response variable that can be explained by the treatment variables) of 0.13, were relatively weak.

[4] ICSEA is the Index of Community Socio-Educational Advantage, which provides an indication of the socio-educational backgrounds of students attending a school.

[5] Mean point was 0.143 on a scale from 0.0 to 1.0.

significant (0.143) increase in average STEM grades. However, looking at how much the attendance rate affected the *change* between Year 9 and 10 grades, the effect was no longer significant.

Conclusions

Despite the limitations of the dataset, the three sets of analyses yielded some noteworthy results, including largely confirming previous analytics research (such as Ismail & Kalsom, 2023; Jeffries, Curtis, & Conner, 2018) on the importance of gender, academic performance, and parental characteristics as predictors of STEM academic outcomes and pursuing STEM education pathways.

Acknowledgements

Thanks to the Parramatta Catholic Education Diocese for providing the dataset, the support of the team at the Diocese (particularly former staff members Rob Wilkins and Raju Varanasi) and to Ian Oppermann (Chair of the Generation STEM Consultative Council) for providing invaluable advice throughout the process. Also, thanks to Roman Marchand Matus (UTS Human Technology Institute, and formerly CSIRO) for providing useful input into the Bayesian network models.

References

Ismail, N. & Kalsom Usof, U. (2023). A systematic literature review: Recent techniques of predicting STEM stream students. *Computers and Education: Artificial Intelligence*, 5. https://doi.org/10.1016/j.caeai.2023.100141

Jeffries, D., Curtis, D.D., & Conner, L.N. (2018). Student factors influencing STEM subject choice in Year 12: a structural equation model using PISA/LSAY data. *International Journal of Science and Mathematics, 18,* 441-461. https://doi.org/10.1007/s10763-019-09972-5

Suggested citation: Banks, C.[1], Gamage, C.[2], Inguruwattage, V.[3], Pinto Moreira, C. [4], Steinburg, D.[1], & Zhao, H.[1] (2024). *Generation STEM: education data analytics project summary*. CSIRO.

Affiliations: [1] CSIRO [2] eWater (formerly CSIRO) [3] Daitum (formerly CSIRO) [3] University of Technology Sydney (formerly CSIRO)