Business Report

Early Intervention Analytics: Predicting Academic Underperformance in Primary Education

Ba Huy Hoang Le

Executive Summary

This report aims to address the sharp rise in students at risk of underperforming in their Year 3 NAPLAN writing test, which has increased by 87.7% from 2016 to 2021.

Using the data collected by 40 schools across Australia and provided by Data2Intel, this has found the following:

1. SES background decreased slightly from Year 1 (102.94) to Year 2 (102.12).
2. Students who perform well in literacy generally do well in numeracy and are less likely at risk of underperforming.
3. Better vocabulary skills in Year 1 lower the risk of underperforming in the Year 3 writing test.
4. Non-disabled students are less likely be at risk of underperforming compared to cognitively disabled students.

The report concludes with the following recommendations to benefit many stakeholders: students, schools, parents, and teachers.

- Use of logistic regression model in this case is suitable and reliable for future prediction
- Students' cohort to be clustered into 2 groups: High literacy performance and Low literacy performance
- Larger data size and inclusion of 2019 data will add more value to the analysis.
- More relevant attributes should be explored for a better cluster profile.

Introduction

Need

The number of students at risk of underperforming in the Year 3 NAPLAN writing test increased by 87.7% from 2016 to 2021, while number of not-at-risk students only increased 13.9% overall after a period of growth by 2020 (1.8 times increase).

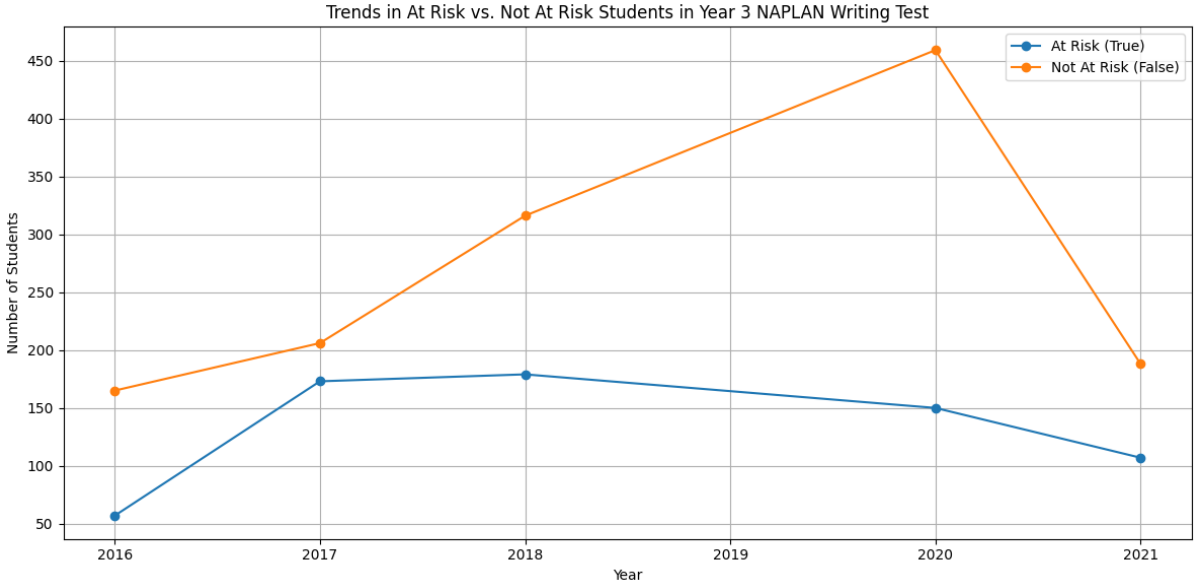

Figure 1. Changes in Year3_At_Risk from 2016 to 2021

Hence, the rising number of students at risk of underperforming must be addressed while the number of well-performing students must be increased.

Changes

With targeted and early intervention, we propose decreasing the percentage of students at risk and increasing the percentage of students not at risk by 5% simultaneously each year using data from 2018 (pre-COVID) as a baseline.

Context

Australia has one of the best education systems in the world (Long, 2023). It starts with the primary education system, in which the curriculum strongly emphasises a balanced literacy approach that includes phonics, comprehension, vocabulary, and writing skills. Students are taught and encouraged to engage with various study approaches in their primary school year, from guided reading, shared reading, and writing workshops to building a strong foundation in reading and writing as they progress in their academic journey (Victoria Department of Education, 2023). However, issues persist, including gaps in literacy outcomes among demographic groupings. Students from low

socioeconomic origins and Indigenous students are more likely to underperform (Australian Institute of Health and Welfare [AIHW], 2023).

Solution

Logistic regression and a K-NN classifier can be used to identify and predict students at risk of writing underperformance based on their literacy and numeracy-oriented assessments.
Apply k-mean clustering to create cluster profiles for targeted intervention and support programs for needy students.

- Students who contribute to the dataset thanks to their test results
- Schools that gather and provide student test scores and their background and characteristics
- Data2Intel and business analysts who analyse data, develop model and provide actionable insights for schools to develop and create plans to support students

Value

- Students can improve their test results, hence their academic results.
- Schools can improve ranking and possibly seek more funding thanks to the increase ranking.
- Data2Intel can access valuable data that can be used in niche domains like the education sector and provide value-added services to benefit schools across Victoria.

Exploratory Data Analysis

1. SES Backgrounds of the Students in Year 1 and Year 2

Figure 2 shows a slight decline in SES from Year 1 (102.94) to Year 2 (102.12). Scores remain above the 2018 national averages of Catholic (100) and independent (102) schools. Both years have similar ranges (42), medians (101), and standard deviations of 9.39 and 9.15, respectively. There are no outliers, indicating a stable socioeconomic profile.

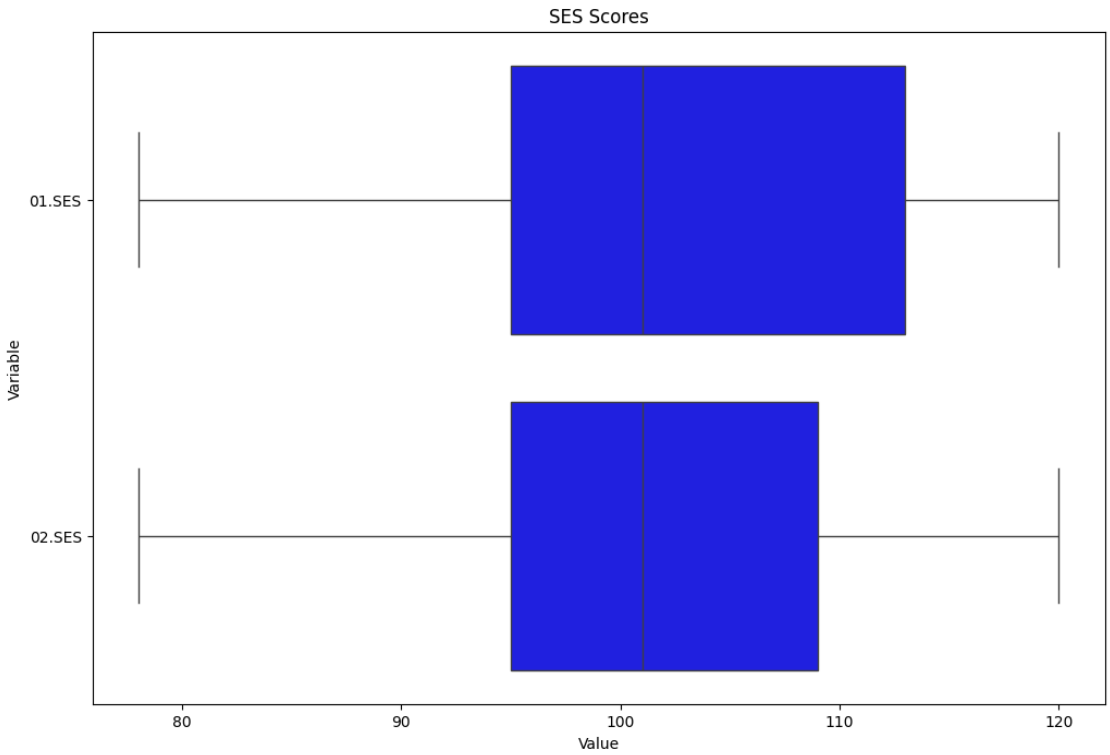

Figure 2. SES Background of Year 1 and Year 2 Students.

However, Figure 3 highlights a significant drop in SES for Year 2 students from 2020 to 2021, with females decreasing from 103.28 to 97.54 and males from 103.33 to 96.58.

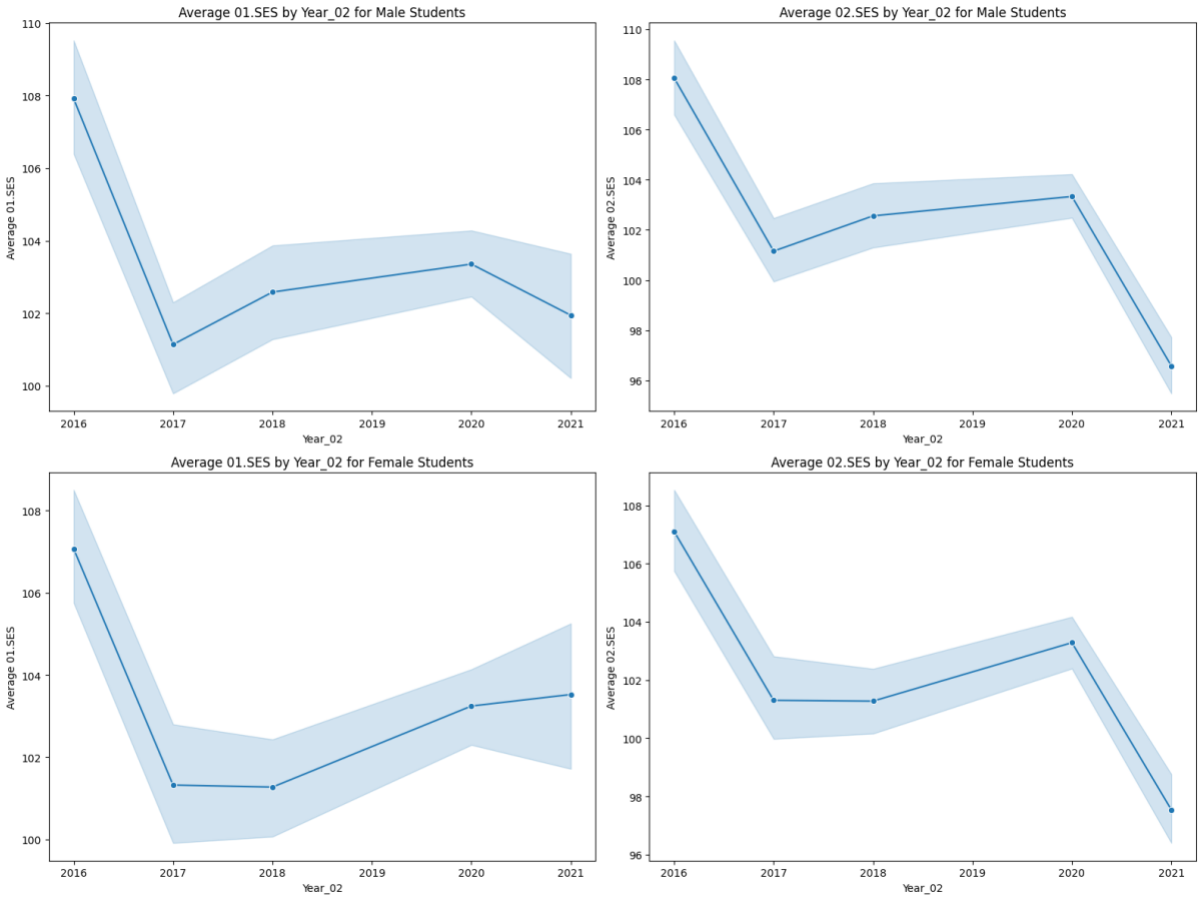

Figure 3. SES Background of Year 1 and 2 Male and Female Students (2016-2021)

Additionally, Figure 4 shows that the SES score tends to be lower as the number of siblings in a family increases.

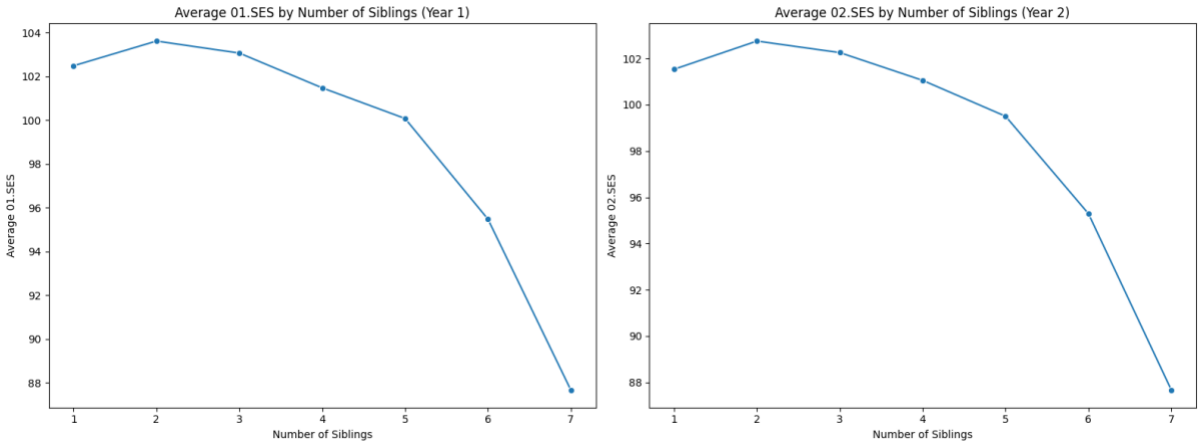

Figure 4. SES Background of Year 1 and Year 2 Students by Number of Siblings

2. Examining Students' Reading Skills

a. Burt Reading Scores

Figure 5 illustrates a substantial improvement in Year 1 students' Burt Reading Scores. The median score was 21 initially and increased to 33.

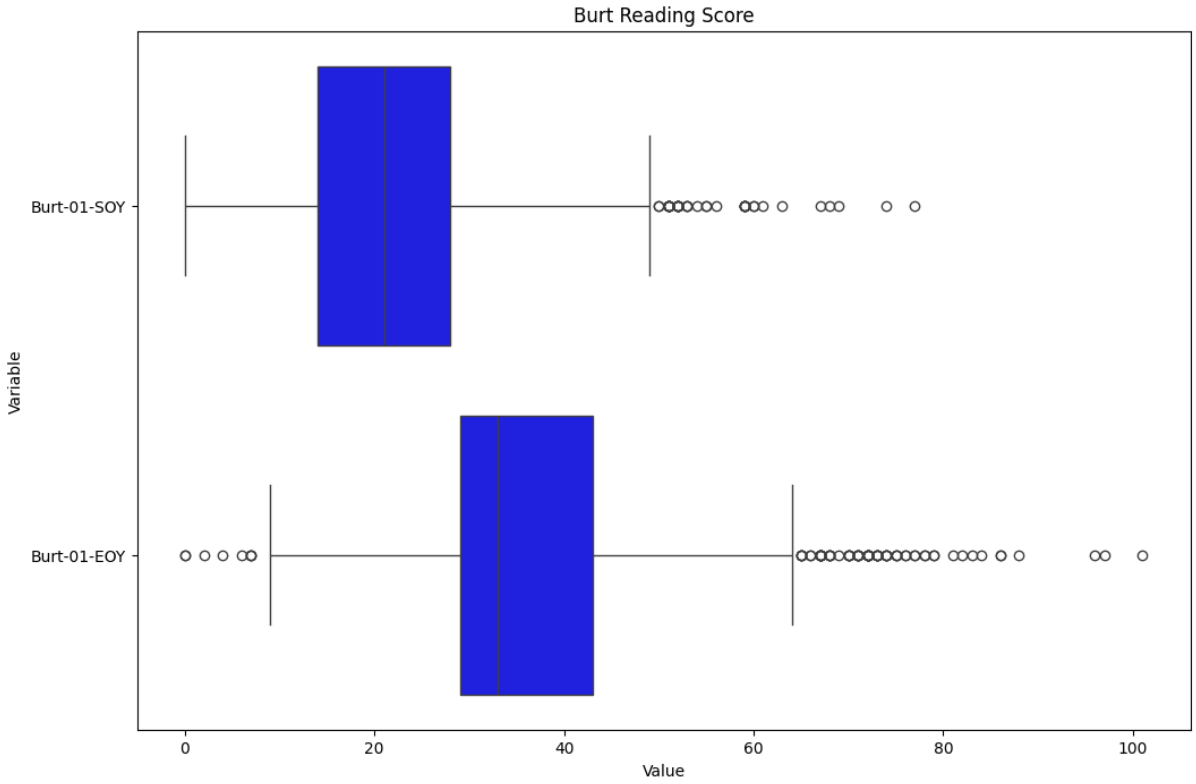

Figure 5. Burt Word Test Scores for Year 1 Students at the Start of Year (SOY) and End of Year (EOY)

b. Clay Reading Scores

Figure 6 shows that average clay reading scores significantly rise from 2.65 to 6.96. However, the increase in standard deviation from 3.59 to 5.88 indicates that some students' gaps in progress widened.

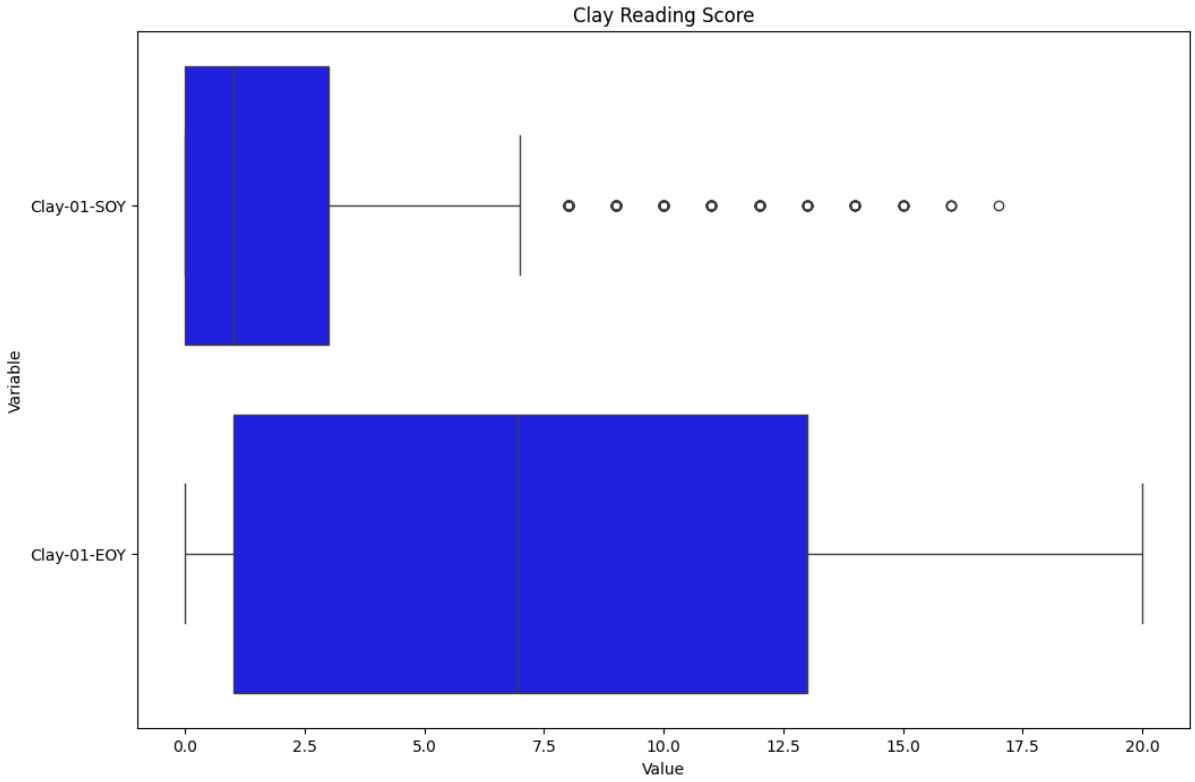

Figure 6. Clay Word Reading Test Scores for Year 1 Students at SOY and EOY.

c. Text Reading Scores

Figure 7 shows that the text reading score for Year 1 consistently increased from 10.7 at SOY to 21.13 at EOY. The decrease in standard deviation from 6.08 at SOY to 4.58 at EOY indicates that student performance has become more consistent throughout the year.

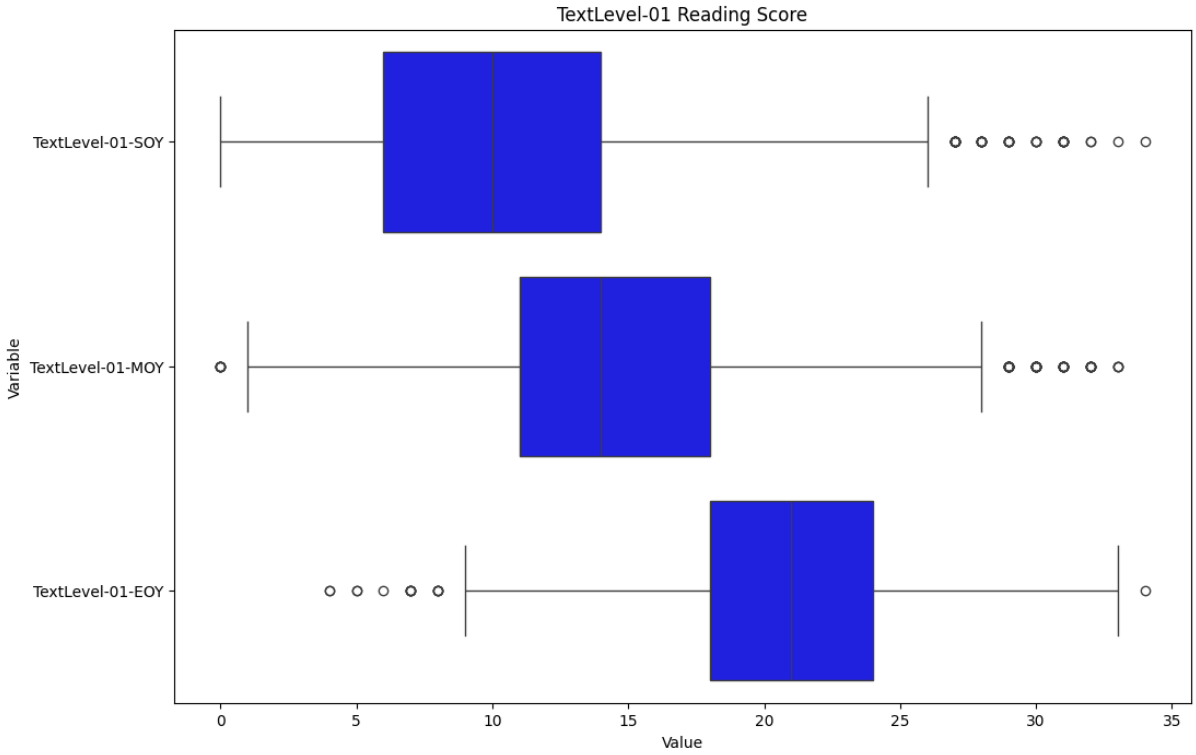

Figure 7. Text reading score for Year 1 students SOY-MOY-EOY

Similarly, Figure 8 shows that the text reading score for Year 2 students consistently improves from 21.79 at SOY to 27 at EOY. The decrease in standard deviation from 5.22 at SOY to 3.77 at EOY indicates that student performance has become more consistent over the year.

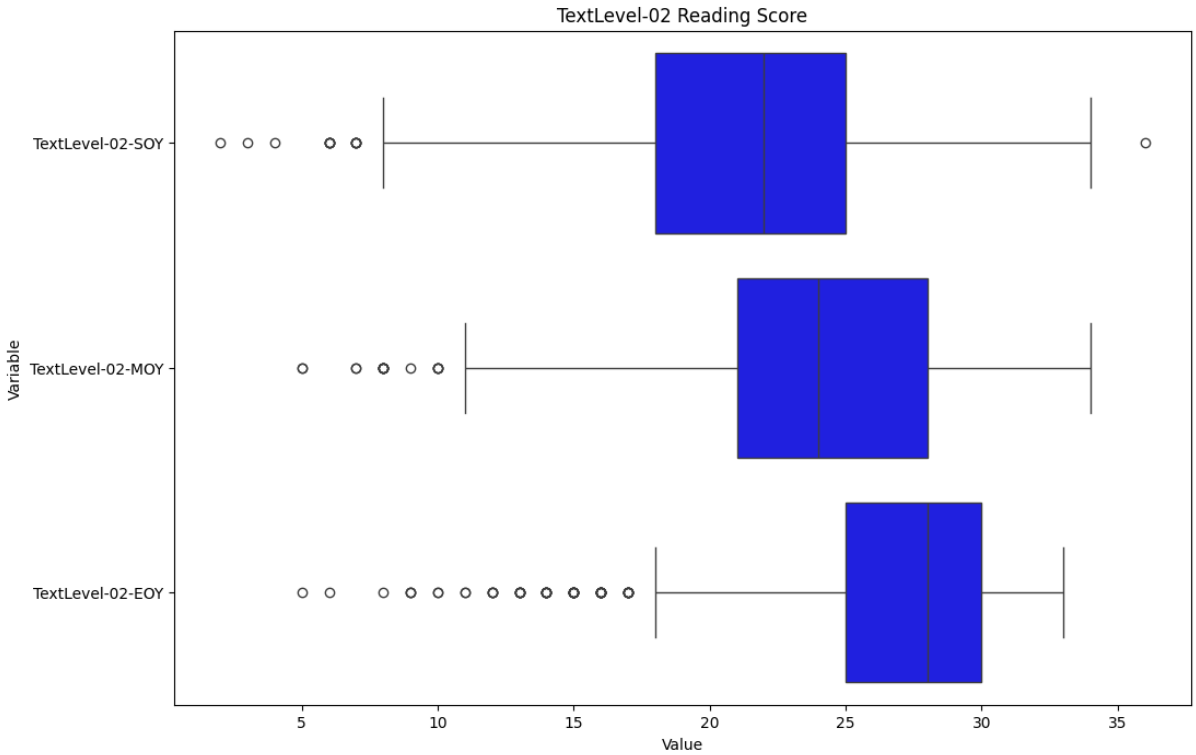

Figure 8. Text reading score for Year 2 students SOY-MOY-EOY

Overall, Year 1 students show a higher percentage increase in their average score, rising by 97.4%, compared to Year 2 students, who show a 23.9% increase.

d. Comparison

As shown by Figure 9, a common trend among all three tests is that students' average scores reduced from 2016 to 2021. The Clay test shows the most significant reduction, with a decrease of 35.5%.

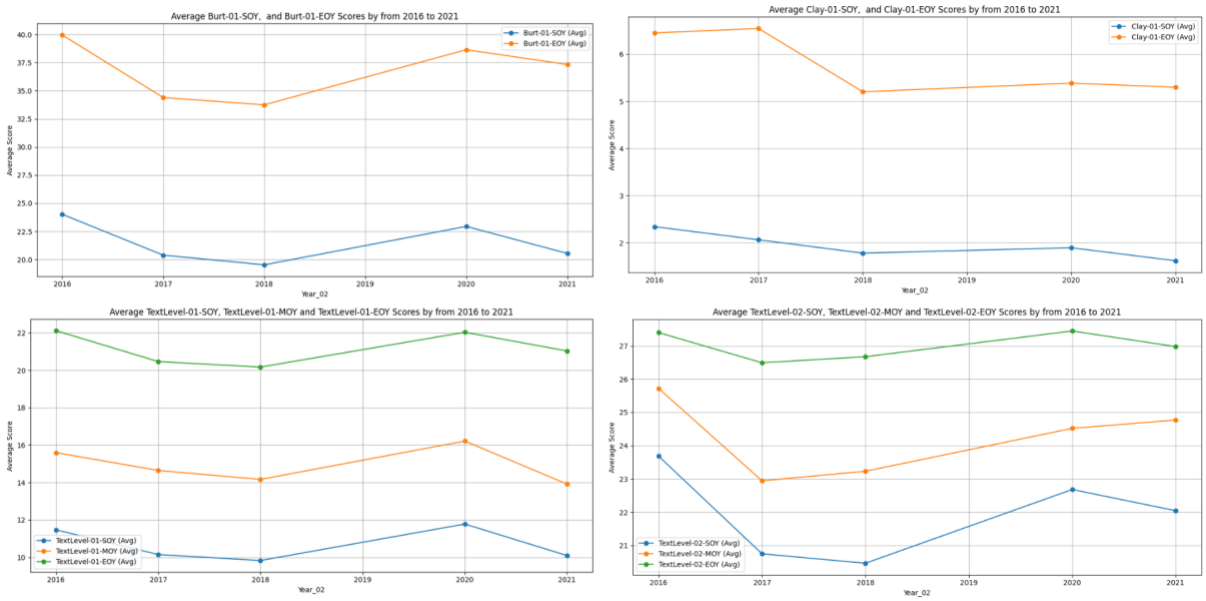

Figure 9. Changes in average reading assessments score from 2016 to 2021

3. Examining Students' Writing Skills at the Start of Year 1

The histogram from Figure 10 shows that the average writing vocabulary score for Year 1 students is approximately 22.0. Based on the boxplot, 39 results were considered outliers. The score range is also significant from 0 to 95.

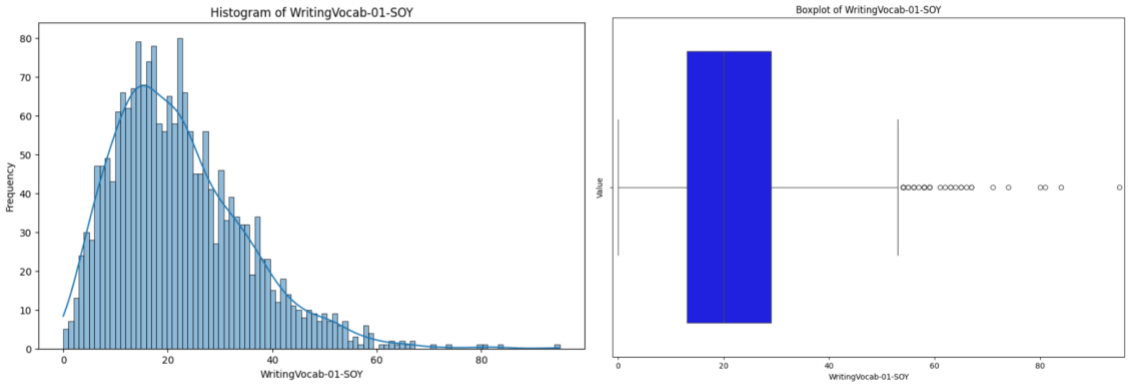

Figure 10. Histogram and Box-plot of WritingVocab for Year 1 student

Similar to other tests, Figure 11 shows a downward trend from 25 in 2016 to 22.8 in 2021.

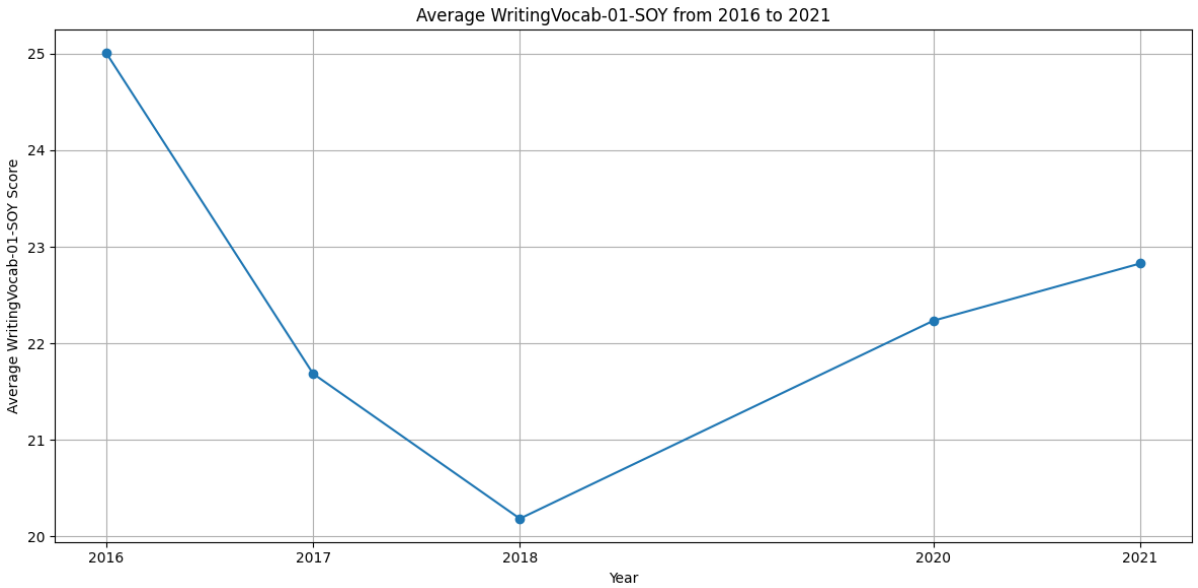

Figure 11. Changes in WritingVocab for Year 1 students from 2016 to 2021

There is a relationship between WritingVocab-01-SOY and Year3_Writing_At_Risk. Figure 12 shows that two-thirds of students classified as not-at-risk have higher writing

vocabulary scores in Year 1 (25.1) than their at-risk peers (15.8).

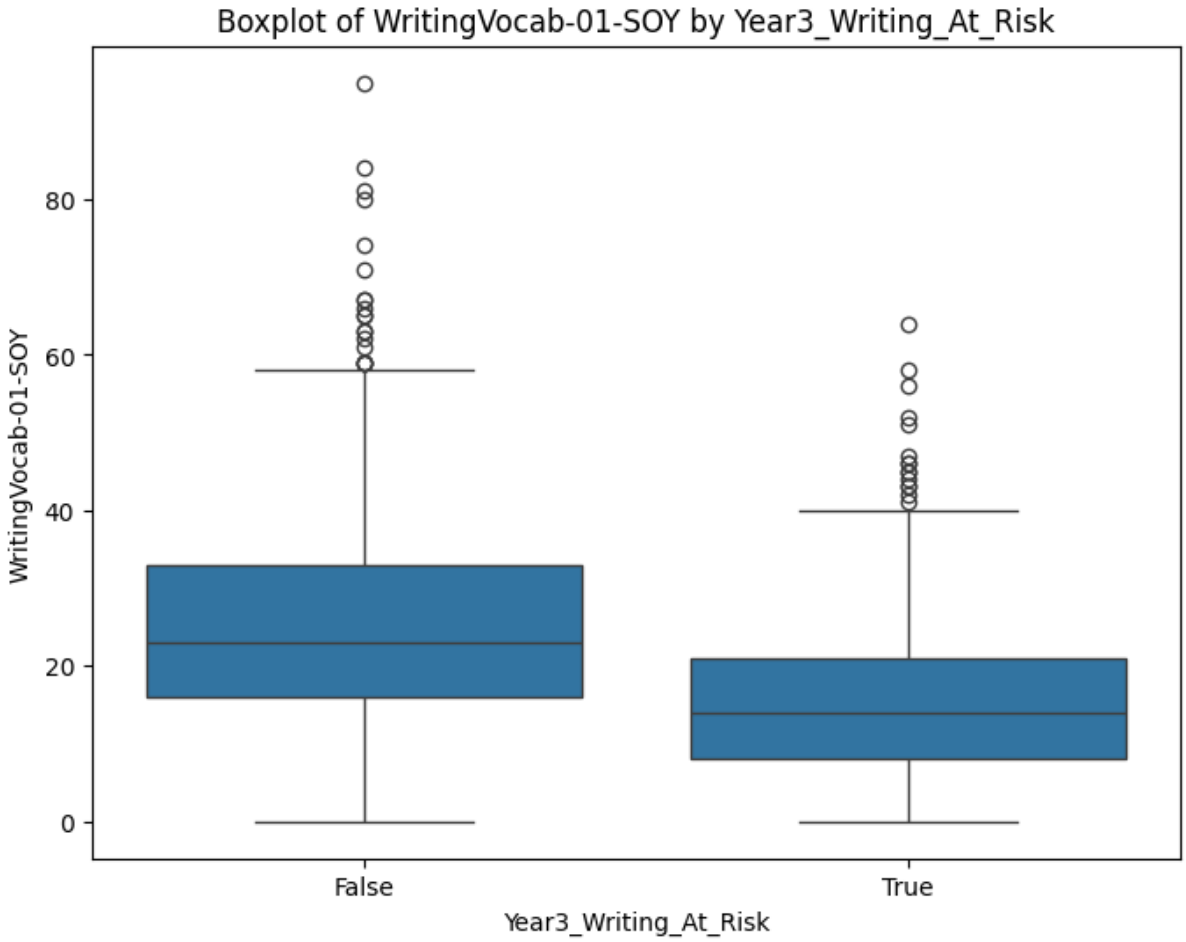

Figure 12. Boxplot of students' writing skills in Year 1 against Year3_Writing_At_Risk

Higher Year 1 writing vocabulary correlates with lower risk in Year 3 writing, confirmed by the negative correlation value of -0.35 (Figure 13).

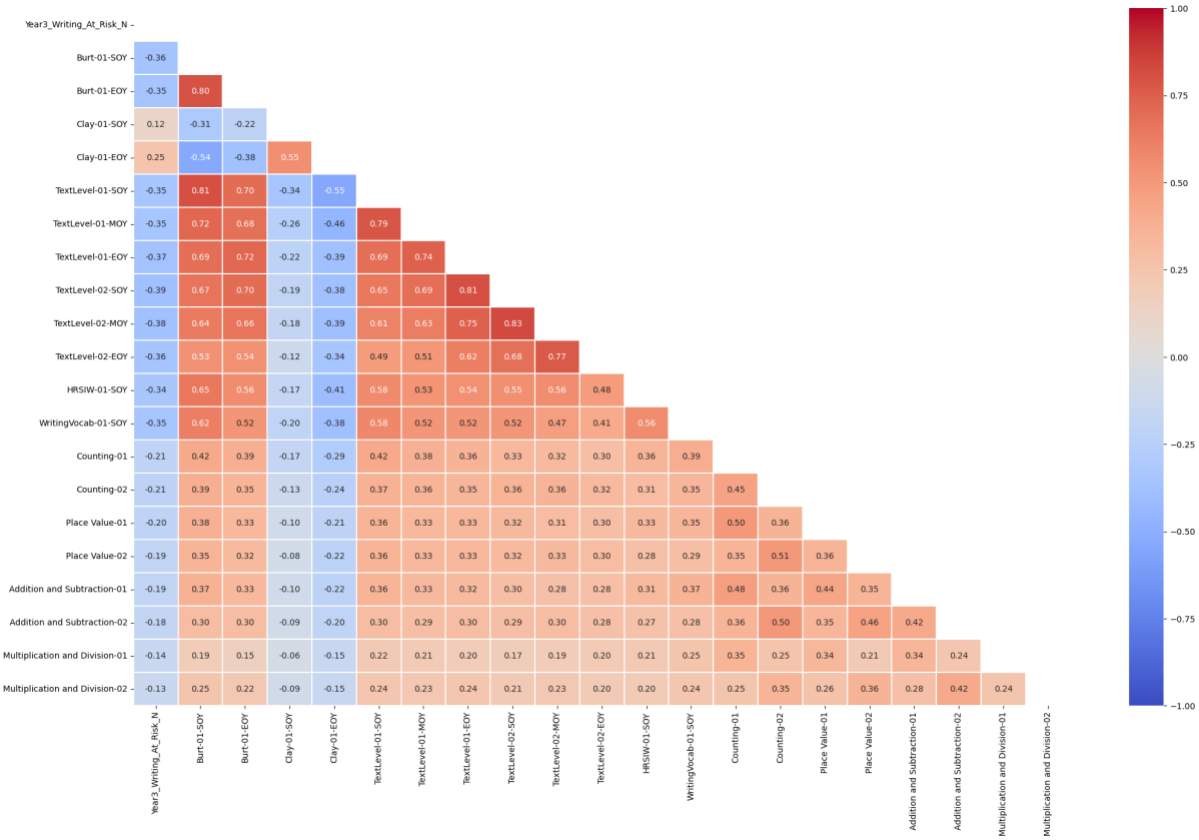

Figure 13. Heatmap of literacy-oriented and numeracy-oriented assessments with Year3_Writing_At_Risk

4. Examining the Relationship of Students' Literacy and Numeracy Skills, and Their Relationships with Year3_Writing_At_Risk

Figure 13 shows that students' literacy and numeracy skills are positively correlated (0.2-0.4), except for the Clay Word Reading test. The highest correlations include Counting-01 with Burt-01-SOY (0.42) and TextLevel-01-SOY (0.42). This suggests that students who perform well in literacy also do well in numeracy.

Figure 13 shows negative correlations between Year3_Writing_At_Risk and most numeracy and literacy assessments, except for the Clay reading test. Higher scores on these tests lower the risk of underperforming in Year 3 writing. Text reading tests have the strongest correlations, ranging from -0.35 to -0.39, while numeracy assessments show weaker correlations, ranging from -0.13 to -0.21.

5. Examining Students' Disability Status and the Relationships with Year3_Writing_At_Risk

Figure 14 illustrates that 69% of students are not disabled, while 31% have some form of disability. Among those with disabilities, the bar chart in Figure 14 reveals that the majority are cognitively disabled, representing 75.8% of disabled students (469 students). In contrast, only 1.6% of disabled students (10 students) have a sensory disability.

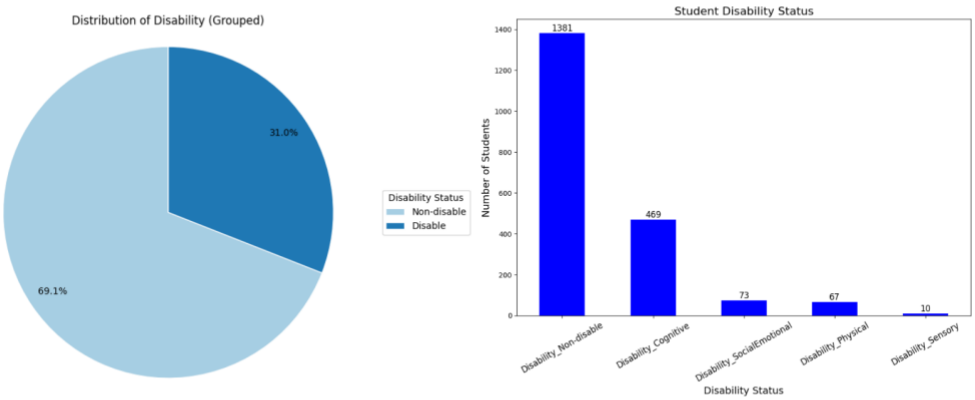

Figure 14. Student disability status

Based on Figure 15, only Disability_Cognitive and Disability_Non-disable correlate with Year3_Writing_At_Risk. Disability_Cognitive has a weak positive correlation (0.25), indicating cognitively disabled students are at a higher risk in Year 3 writing. Disability_Non-disable has a weak negative correlation (-0.24), suggesting a lower risk for non-disabled students.

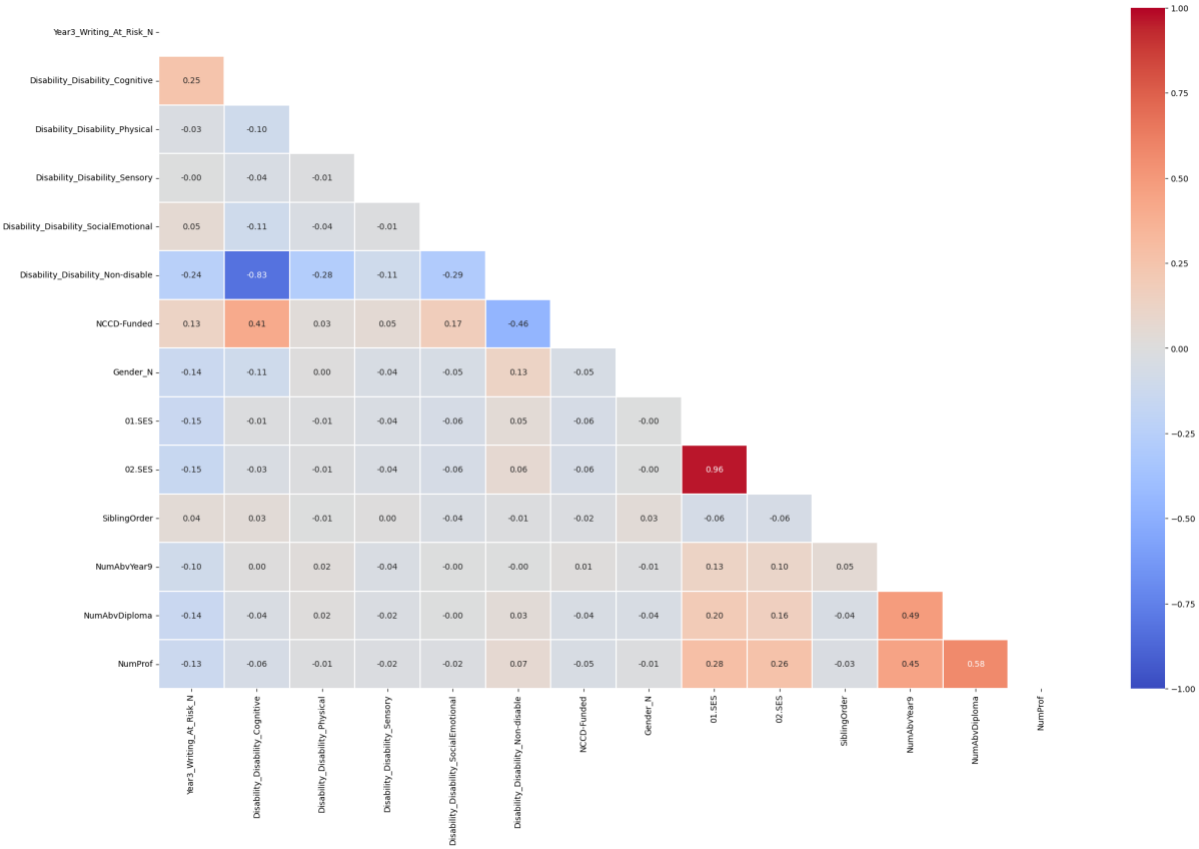

Figure 15. Heatmap of student characteristics and background with Year3_Writing_At_Risk

Proposed Machine Learning Solution

1. Supervised Machine Learning

This report proposes using a Logistic Regression model to classify Year3_Writing_At_Risk. The output is shown in Appendix 1. Figure 16 shows the performance metrics of the post-optimised and validated (k-fold validation, k =10) model.

- Accuracy = 0.730 +/- 0.012: The model has moderate predictive power. It correctly classifies performance status for about 73%.
- Precision for “True” = 0.488. Only 49 out of 100 students are predicted by the model to be “at risk.”
- Recall for “True” = 0.767. The model can correctly identify 76.7% of the students who were actually “at risk.”
- F1 score for “True” = 0.506 +/- 0.028. The moderately high F1 score means the model has a moderate balance between precision and recall.
- AUC = 0.76. The model’s predictive ability in distinguishing between at-risk and not-at-risk students is relatively good.

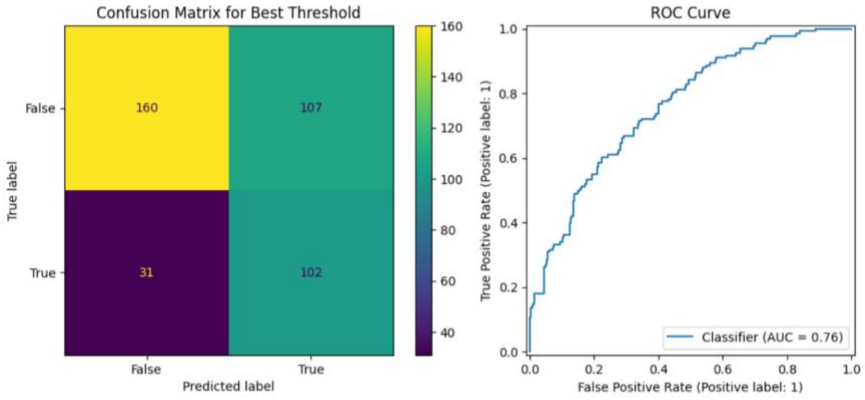

Figure 16. Confusion matrix and ROC curve

Pros	Cons
Quick training time (less than 1 second for the current dataset) and easy to interpret (thanks to the equation output - Appendix 1)	Low precision rate
Relatively high recall rate	Dataset is small, model results might not be a good representative if applied on large set of Australian students. Overfitting is potential as dataset is small
Relativey high accuracy rate	Model is sensitive with outliers

Table 1. Pros and cons of the Logistic Regression model

2. Unsupervised Machine Learning

The proposed clustering model is k-means with k = 2 (2 clusters), as shown in Figure 2. Perfomance metrics for this model.

- WCSS score = 319.97
- Davies Bouldin index = 0.608
- Silhouette score: 0.652

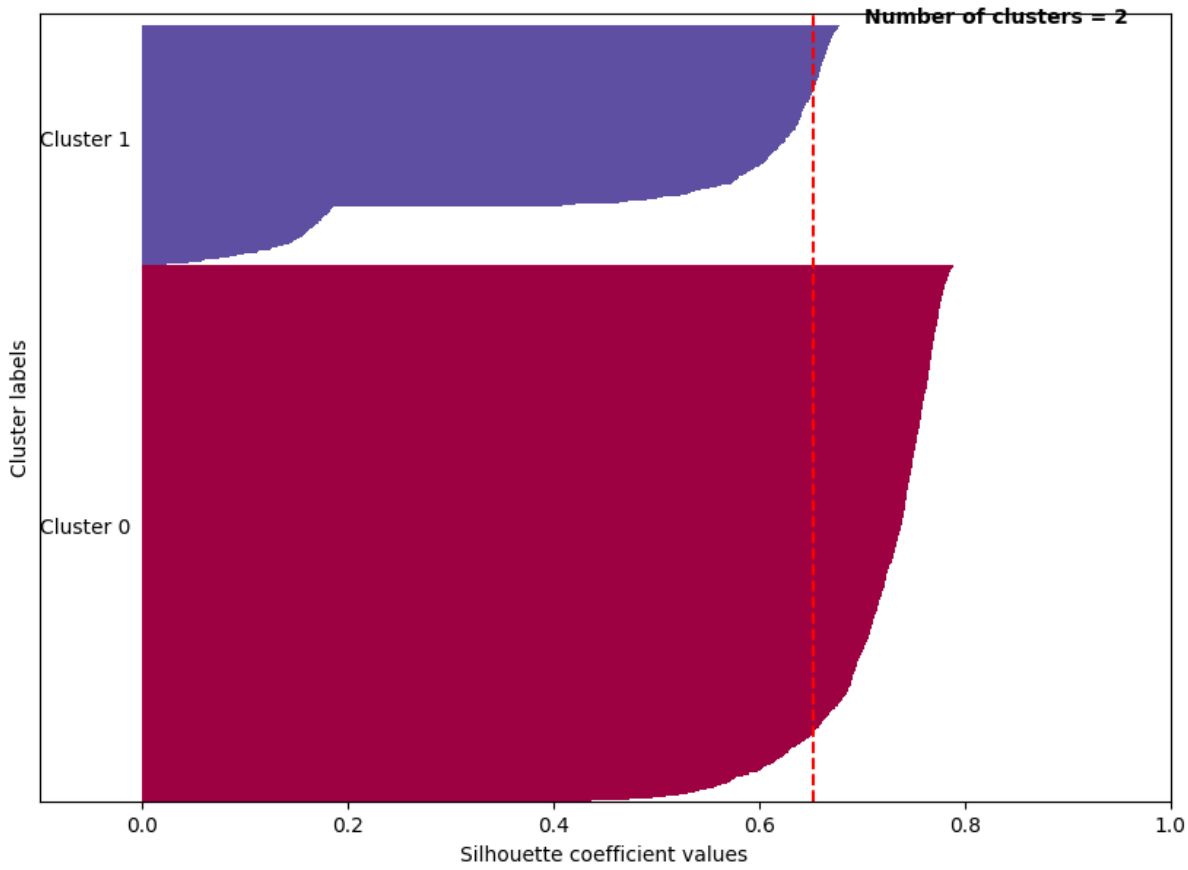

Figure 17. Student Clusters with k = 2

Post-analysis from Figure 18 shows that cluster 0 accounts for 69% and cluster 1 for 31% of the data.

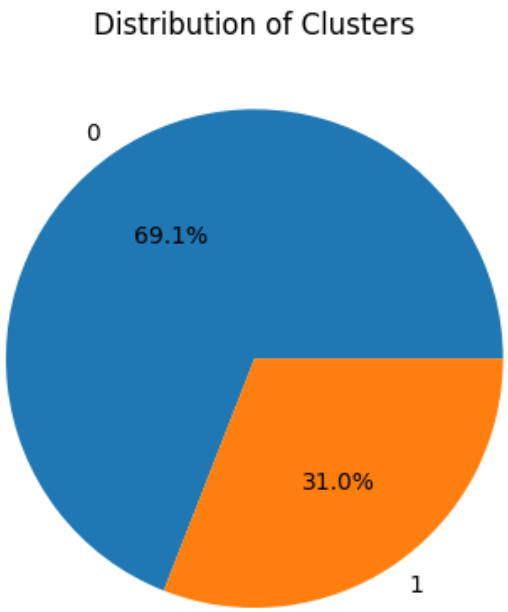

Figure 18. Distribution of clusters

Issues with the current clustering model

- Both clusters are quite similar (Figure 19)

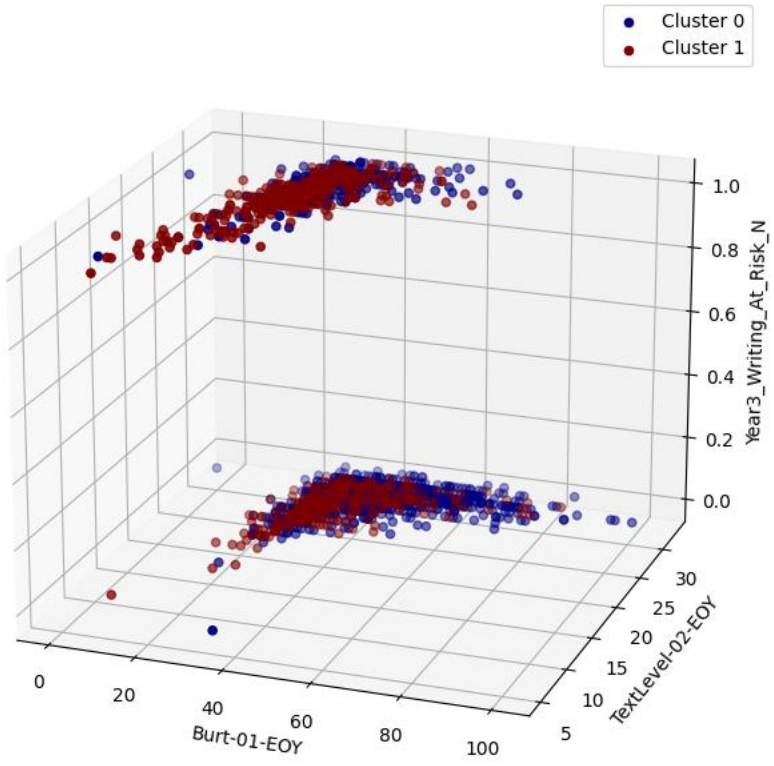

Figure 19. 3-D sketch of Cluster 0 and 1

- Student demographics and characteristics were not considered due to their low correlation coefficient values

Recommendations and Conclusions

The goal is to improve students' Burt reading tests at the end of Year 1 and Text reading tests at the end of Year 2. Assist and instruct students who are showing early signs of difficulties in these areas. Make one-on-one or small group reading sessions a priority. Teachers and parents are encouraged to work together to evaluate progress and ensure students gain the fundamental reading skills required for better writing results while engaging families in encouraging practice at home.

A summary of cluster profiles and support for each cluster is shown in Table 2.

	Cluster 0	Cluster 1
Name	High literacy performance	Low literacy performance
Disability status	No disabled students	Cognitively disabled students
Support	Promoting reading activities and creative writing for students to further improve their skills	Visual aids and other tools to enhance study experience Constant communication between teachers and parents to update and keep track Study group with high performing student group

Table 2. Cluster profile

Benefits for stakeholders are shown in Table 3.

Stakeholders	Benefits
Students	Enhance literacy skills especially in early years of struggle to prevent discouragement Promote better Improve class rank
Teachers	Better understand study to tailor study plan for individual students
Parents	Better communication between teachers and parents Understand student more in how to provide out-of-school support
Schools	Improve school ranking Better resource allocation and reduce waste of human and financial resources

Table 3. Benefits for different stakeholders

Implications for business process and decision-making

- Changes in study programs, especially in reading and writing

- Re-allocation of state and schools resources to provide target support for students in need
- Data-driven insights become the crux of school decision to implement targeted implementation for students in need of support

Further improvement of data

- Larger size of data for more reliable model and better-aligned results with the real-world
- Explore more closely correlated features and more relevant student demographic data for more distinct cluster
- Data in 2019 can add further values for the analysis

Reference

Australian Insitute of Health and Welfare. (2023, December 07). *Education of First Nations people*. https://www.aihw.gov.au/reports/australias-welfare/indigenous-education-and-skills

Department of Education Victoria. (2023, April 19). High impact teaching strategies (HITS). https://www.education.vic.gov.au/school/teachers/teachingresources/practice/improve/Pages/hits.aspx

Long, C. (2023, December 05). Australia is now in the world's top 10 academic performers – but the data paints a complex picture. *ABC News*. https://www.abc.net.au/news/2023-12-05/pisa-international-school-rankings-in-maths-science-reading/103185468

Appendix

Appendix 1 – Logistic regression output

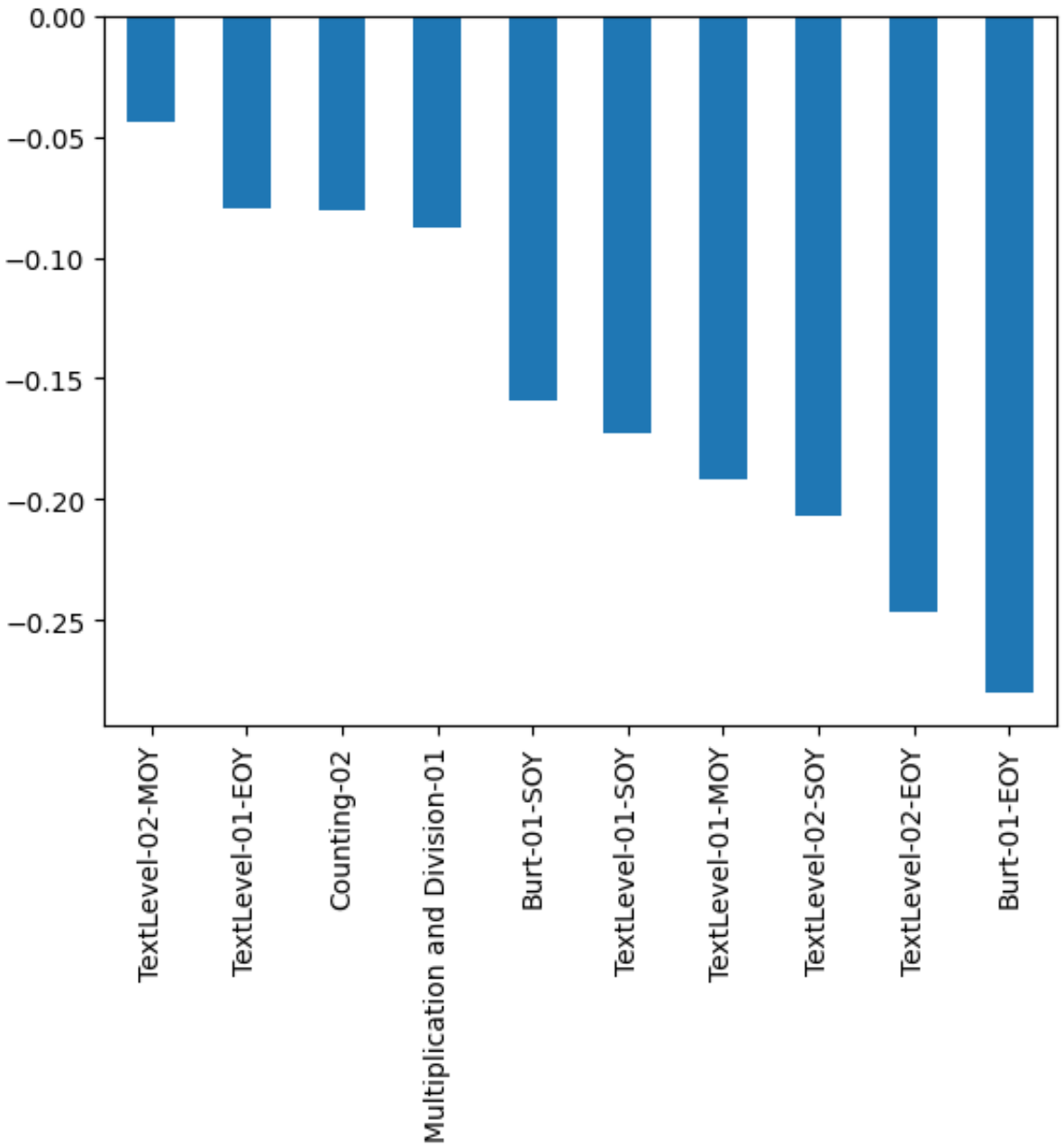

Figure 20. Visualisation of the coefficients

- Literacy-oriented assessments generally have a stronger predictive effect against the target feature than numeracy-oriented assessments.
- Burt-01-EOY (-0.28) and TextLevel-02-EOY (-0.247) have the strongest influence in reducing underperformance risk. For every one-point increase in Burt-01-EOY, the odds of being at risk are reduced by 24.4%. Similarly, the odds of being at risk are reduced by 22.3%, for every one-point increase in TextLevel-02-EOY.
- Counting-02 (-0.080) and Multiplication and Division-01 (-0.088): For every one-point increase in Counting-02, the odds of being at risk are reduced by 7.7%. The odds of

being at risk are reduced by 8.4% for every one-point increase in Multiplication and Division-01.

The equation is shown as

Year_3_Writing_At_Risk = -0.893 + -0.159 * Burt-01-SOY + -0.280 * Burt-01-EOY + -0.173 * TextLevel-01-SOY + -0.192 * TextLevel-01-MOY + -0.080 * TextLevel-01-EOY + -0.206 * TextLevel-02-SOY + -0.044 * TextLevel-02-MOY + -0.247 * TextLevel-02-EOY + -0.080 * Counting-02 + -0.088 * Multiplication and Division-01

Since Burt-01-EOY and TextLevel-02-EOY contribute the most to the classification model, the report recommends focusing on improving the skills in these for better test results and lower chance of being at risk when it comes to Year 3 NAPLAN writing test.