

Analytic Report

# **Early Intervention Analytics: Predicting Academic Underperformance in Primary Education**

Ba Huy Hoang Le

<b>EXECUTIVE SUMMARY</b>	<b>3</b>
<b>INTRODUCTION</b>	<b>4</b>
<b>APPROACH</b>	<b>4</b>
<b>DATA PREPARATION AND EXPLORATORY DATA ANALYSIS (EDA)</b>	<b>5</b>
1. <b>Data Observations and Pre-processing</b>	<b>5</b>
2. <b>Univariate Analysis</b>	<b>6</b>
a.    Year3_Writing_At_Risk	6
b.    Gender	7
c.    Disability Status	8
d.    SES Background	10
e.    Burt Word Test	11
f.    Clay Word Reading	12
g.    Text Reading	13
3. <b>Bivariate and Multivariate Analysis</b>	<b>15</b>
a.    WritingVocab-01-SOY and Year3_Writing_At_Risk	15
b.    Literacy-oriented and Numeracy-oriented vs. Year 3_Writing_At_Risk	15
c.    Disability and Year3_Writing_At_Risk	15
4. <b>Data Conversion</b>	<b>16</b>
<b>MODEL DEVELOPMENT AND EVALUATION</b>	<b>17</b>
1. <b>Logistic Regression</b>	<b>17</b>
a.    Model Justification	17
b.    Features Selection	18
c.    Model Results and Interpretation	18
d.    Model Performance	19
2. <b>K-Nearest Neighbour</b>	<b>20</b>
a.    Model Justification	20
b.    Features Selection	20
c.    Model Development	21
d.    Model Performance	21
e.    Optimising k	22
3. <b>Model Comparison and Selection</b>	<b>22</b>
4. <b>K-means Clustering</b>	<b>23</b>
a.    Model Justification	23

b.	Features Selection	24
c.	Model Development	24
d.	Model Performance	24
e.	Model Optimisation	25
f.	Model Selection	27
g.	Post-Analysis	29
<b>SOLUTION RECOMMENDATION</b>		<b>31</b>
1.	<b>Classification Model</b>	<b>31</b>
2.	<b>Clustering model</b>	<b>32</b>
3.	<b>Future engagements with the client</b>	<b>32</b>
<b>TECHNICAL RECOMMENDATIONS</b>		<b>33</b>
1.	<b>Programming Language</b>	<b>33</b>
2.	<b>Machine Learning Diagram</b>	<b>33</b>
a.	Data Collection	33
b.	Data Pre-processing and Cleaning	33
c.	Exploratory Data Analysis	34
d.	Model Training and Evaluation	34
e.	Model Deployment	34
f.	Model Monitor and Feedback	34
3.	<b>Maintaining Model's Accuracy and Relevance</b>	<b>35</b>
<b>APPENDIX</b>		<b>36</b>
<b>Appendix 1. Heatmap between literacy-oriented and numeracy-oriented assessments with Year3_Writing_At_Risk</b>		<b>36</b>
<b>Appendix 2. The decision to prioritise recall over precision</b>		<b>36</b>

# Executive Summary

This report aims to classify students at risk of underperforming in the Year 3 writing NAPLAN test. By correctly and timely identifying these students, target and early intervention can be provided. Furthermore, by classifying students, schools can allocate resources more effectively to provide adequate tools and support for students.

This report explores various aspects of the dataset and the relationship between literacy-oriented and numeracy-oriented assessments and Year 3\_Writing\_At\_Risk (target features). Most literacy—and numeracy-oriented assessments are moderately correlated with the target features. Two classification models were developed: logistic regression and K-NN. A clustering model using k-means methods was deployed to develop cluster profiles for support programs.

This report recommends using a Logistic regression model to classify students. The classification model is reliable (accuracy:  $0.730 \pm 0.012$  and F1 score  $0.506 \pm 0.028$ ). However, there is room for improvement, especially in combining features. Based on the clustering model, this report recommends grouping the students into two groups based on their Burt test scores at the end of Year 1 and Text reading scores at the end of Year 2. Group 1 is High literacy performance, and Group 2 is Low literacy performance.

Future work includes expanding the dataset and adding more relevant features, such as race and other socioeconomic aspects and behavioural data, to improve the model's prediction ability and create a more distinct cluster profile for more targeted support.

## Introduction

This case study uses data from 40 Australian schools provided by Data2Intel to identify students likely to perform below expectations on the Year 3 NAPLAN writing test. The objective is to create prediction models and clusters to help educators implement targeted early interventions. By identifying these at-risk pupils, schools can better deploy resources and develop support programs that directly improve reading outcomes. The value proposition is the early detection of patterns in student performance, which enables prompt intervention and improves educational outcomes. This is important for data-driven educational decision-making.

## Approach

The machine learning techniques used in this dataset include

- Supervised machine learning: Logistic regression and K-NN. The purpose is to find out the best model for future use in predicting and classifying students who are at risk of underperforming in writing in Year 3
- Unsupervised machine learning: Clustering model. The purpose is to explore possible clusters of students to customise targeted solutions in supporting students at risk of underperforming in writing Year 3.

# Data Preparation and Exploratory Data Analysis (EDA)

## 1. Data Observations and Pre-processing

Data2Intel provides 2000 students across 40 schools. Each student is represented with detailed information concerning their performance in Year 1 and 2, including five literacy-oriented and four numeracy-oriented assessments. The dataset includes each student's personal and family background and school characteristics. There are no missing values for all 34 columns.

The dataset contains the following Data type

- One Boolean: Year3\_Writing\_At\_Risk (Categorical, binary: True/False)
- One Float64: Kinder\_Age (Continuous numerical variable with decimal)
- 29 int64: School characteristics, Family characteristics, Student assessments  
(Discrete numerical variables)
- Three Object: Student ID (combination of letters and numbers), Gender and  
Disability (Categorical variables)

Generally, the dataset has no major data errors. However, a few attributes have some errors:

- NumbAbvYear9: A value of 3 is recorded for one student (SN84143480), which is invalid as this attribute should only contain 0, 1, and 2. This value is replaced by the mean value of 1.56, rounded to 2.
- Clay-01-SOY, Clay-01-EOY, and TextLevel-01-SOY: some students receive negative scores, which are replaced with 0.

The target variable in this dataset is Year3\_Writing\_At\_Risk

## 2. Univariate Analysis

### a. Year3\_Writing\_At\_Risk

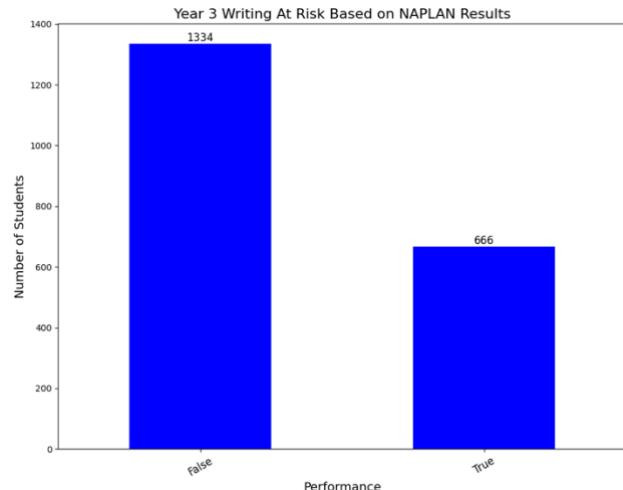


Figure 1 illustrates that 1334 (66.7%) students are not at risk of writing underperformance.

Figure 1. Distribution of at-risk and non-at-risk of underperforming in Year 3 writing NAPLAN test

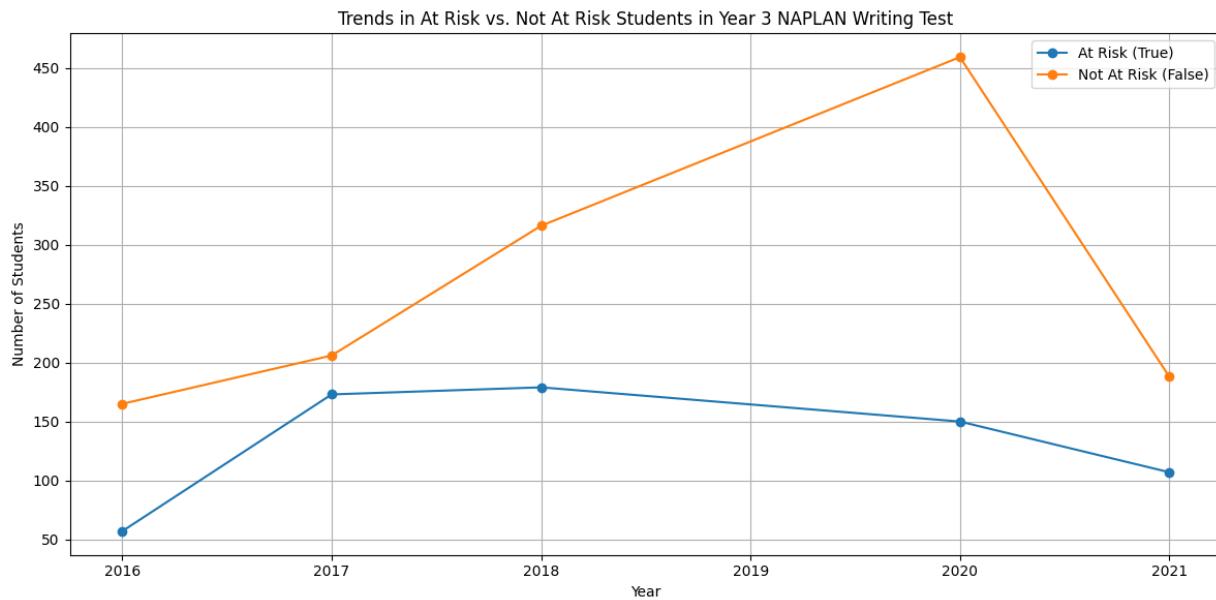


Figure 2. Changes of at-risk and non-at-risk of underperforming in Year 3 writing NAPLAN test from 2016 to 2021

Figure 2 shows a continuous increase from 2016 to 2020 before dropping significantly from 459 to 188 in 2021. Despite this decline, there is an overall increase in these students. However, students at risk have risen considerably quicker, by 1.90 times, compared to the 1.13 times growth of students who are not at risk during the same period.

### b. Gender

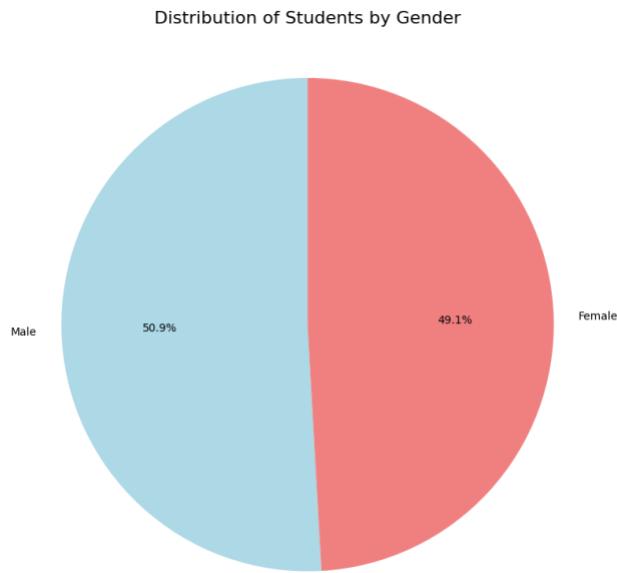


Figure 3 shows that Male and Female students are approximately equal, 50.9% vs. 49.1%, respectively.

Figure 3. Proportion of Male and Female students

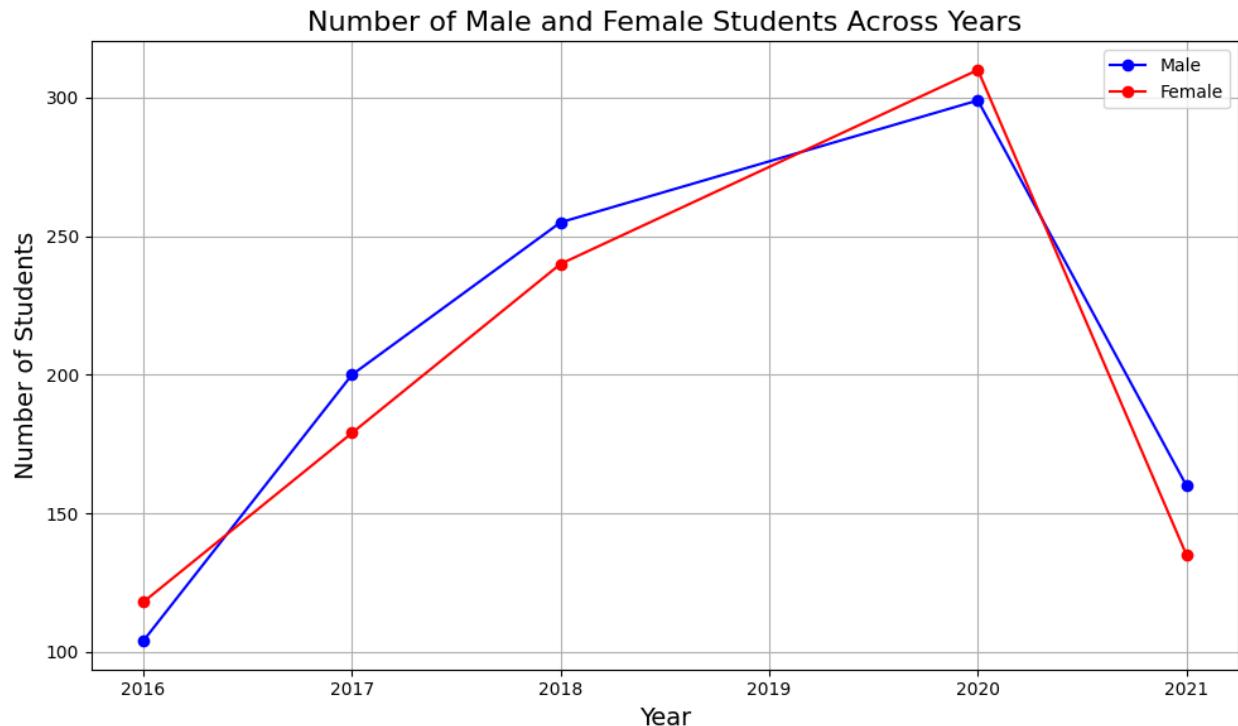


Figure 4. Changes in student gender proportion from 2016 to 2021

From Figure 4, the proportion gap between Male and Female students fluctuated dramatically. The proportions of male students were higher than those of female students in 2017, 2018, and 2021.

### c. Disability Status

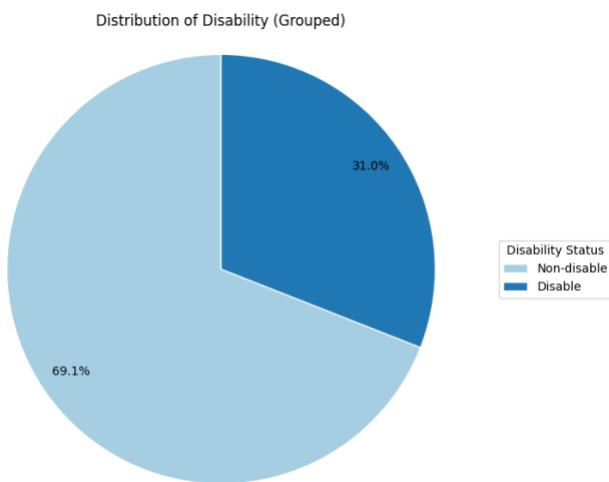


Figure 5 shows that nearly 70% of students have no disabilities, while 31% have some

Figure 5. Distribution of disabled and non-disabled students

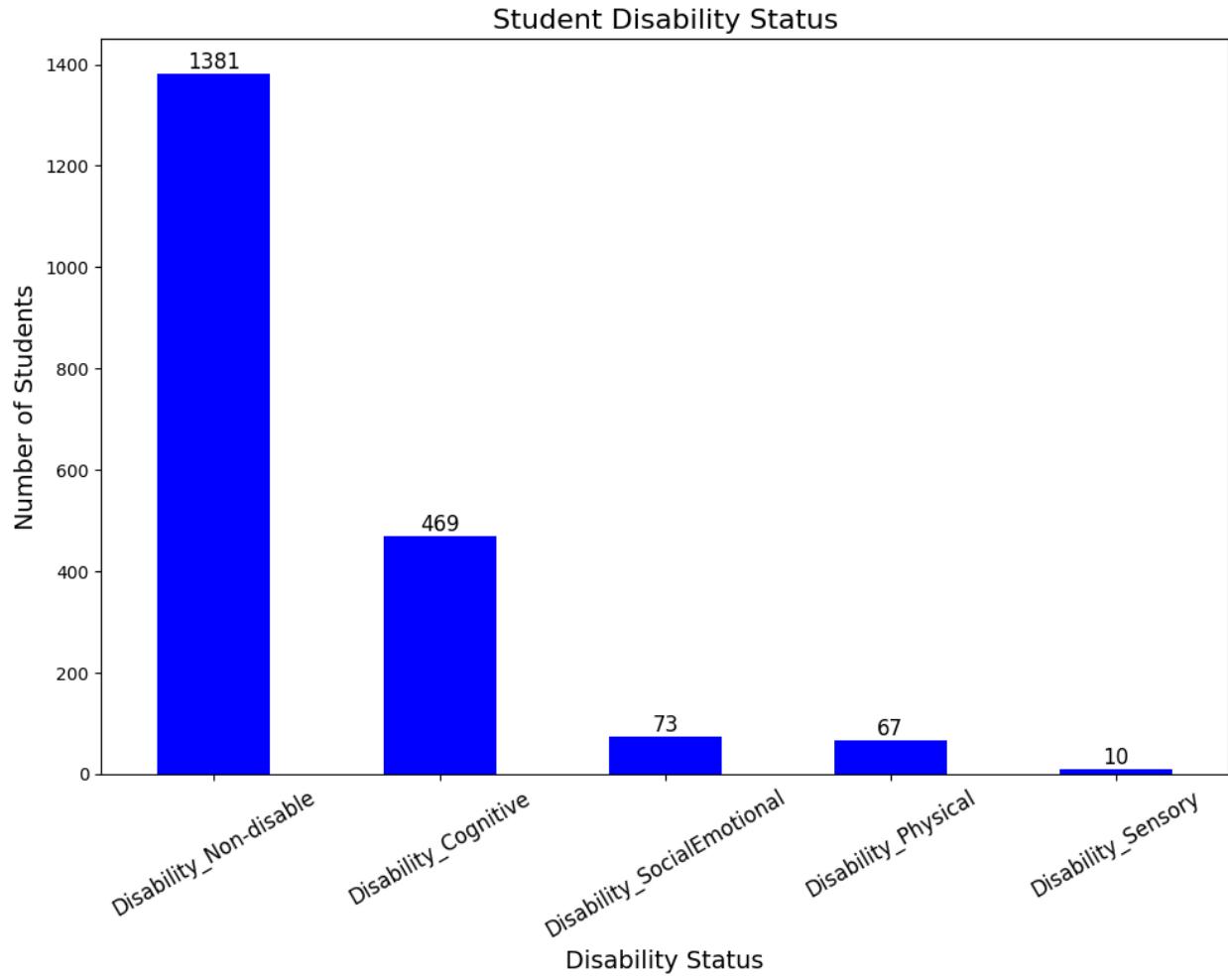


Figure 6. Student disability status

Figure 6 shows that students with disabilities are cognitively disabled, accounting for 75.8% (469 students). In comparison, only 1.6% of disabled students (10 students) have a sensory disability.

#### d. SES Background

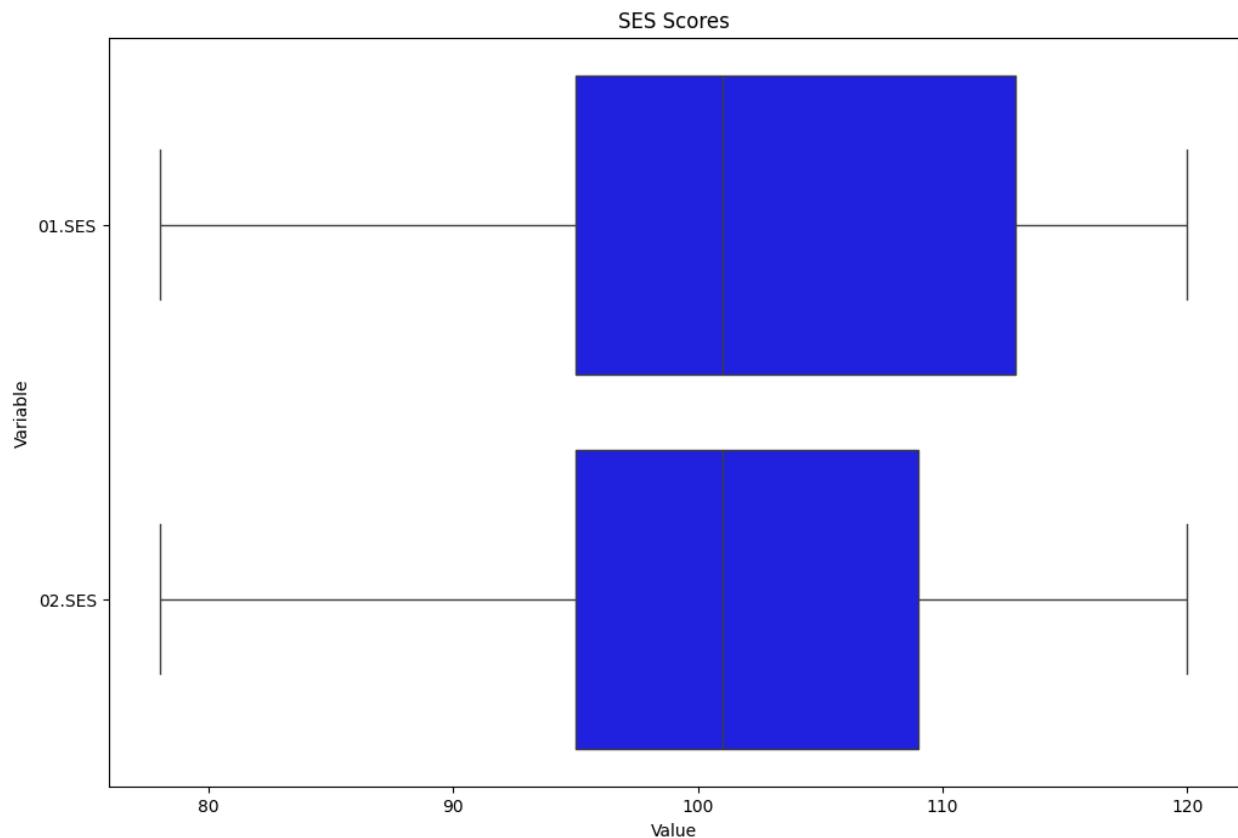


Figure 7. SES status of the school where the student attends in Year 1 and 2

Figure 7 shows the SES data in Year 1 (01.SES) has a mean of 102.94, a range of 42, and slight positive skewness, as the mean is higher than the median of 101. By Year 2 (02.SES), the mean SES dropped to 102.12, although the range remains at 42. The median also remains at 101, with the same slight positive skewness. The standard deviations at SOY and EOY are quite similar, 9.39 and 9.15, respectively. No records are considered outliers in any year, indicating a steady socioeconomic profile with a slight reduction with time.

### e. Burt Word Test

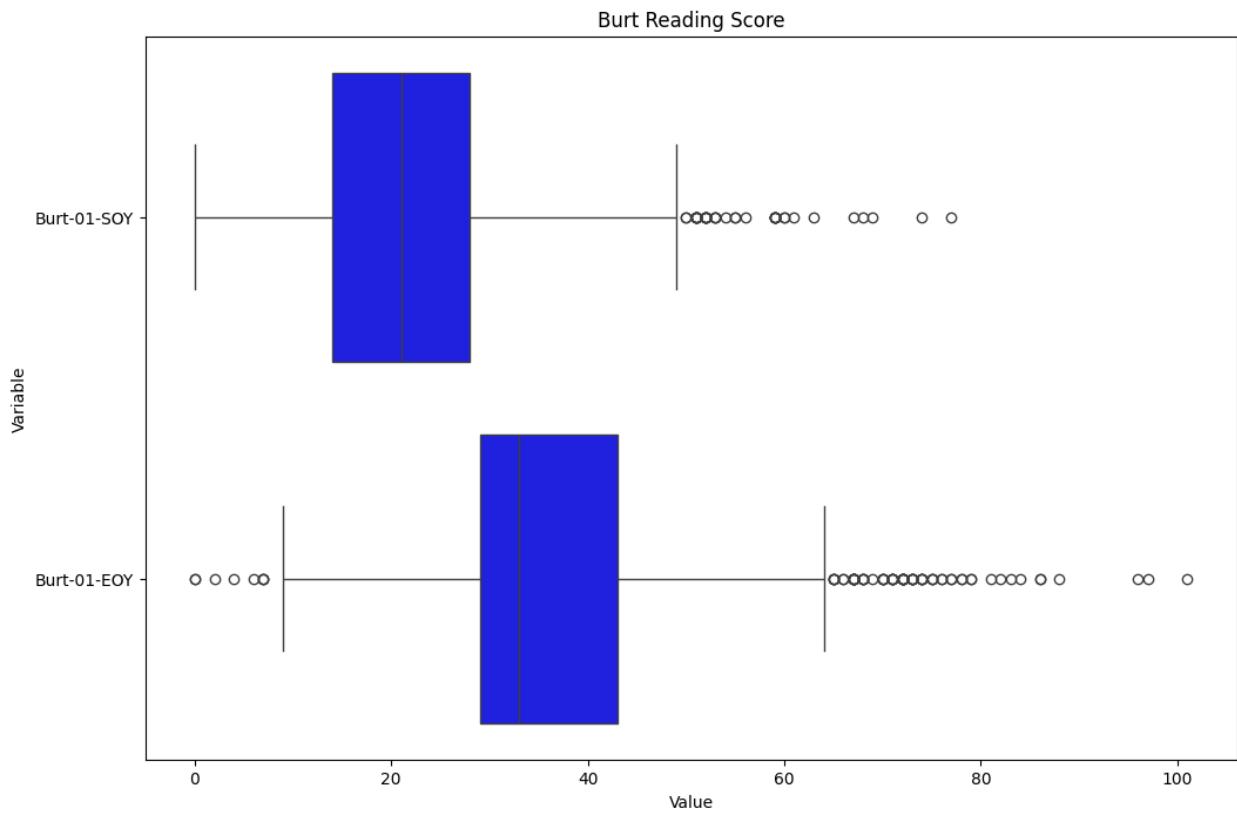


Figure 8. Box plot of Burt's reading score at the start of year (SOY) and end of year (EOY) for  
Year 1 students

From Figure 8, the mean grew dramatically from 21.38 at the SOY to 36.57 at EOY. The standard deviation also increased from 11.18 to 12.26, indicating more variation in performance as the year went on. There are more substantial outliers at EOY, with 74 versus 39 at SOY.

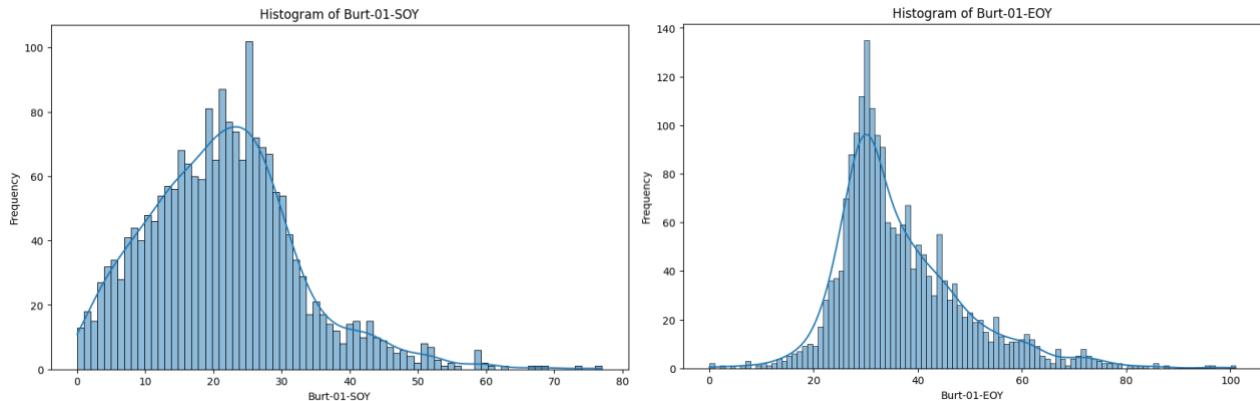


Figure 9. Histograms of Burt's reading score at the start of year (SOY) and end of year (EOY) for Year 1 students

The data for Burt-01-SOY and Burt-01-EOY are positively skewed, suggesting that many students scored lower than the average, but a few top performers drove the mean higher, as shown in Figure 9.

#### f. Clay Word Reading

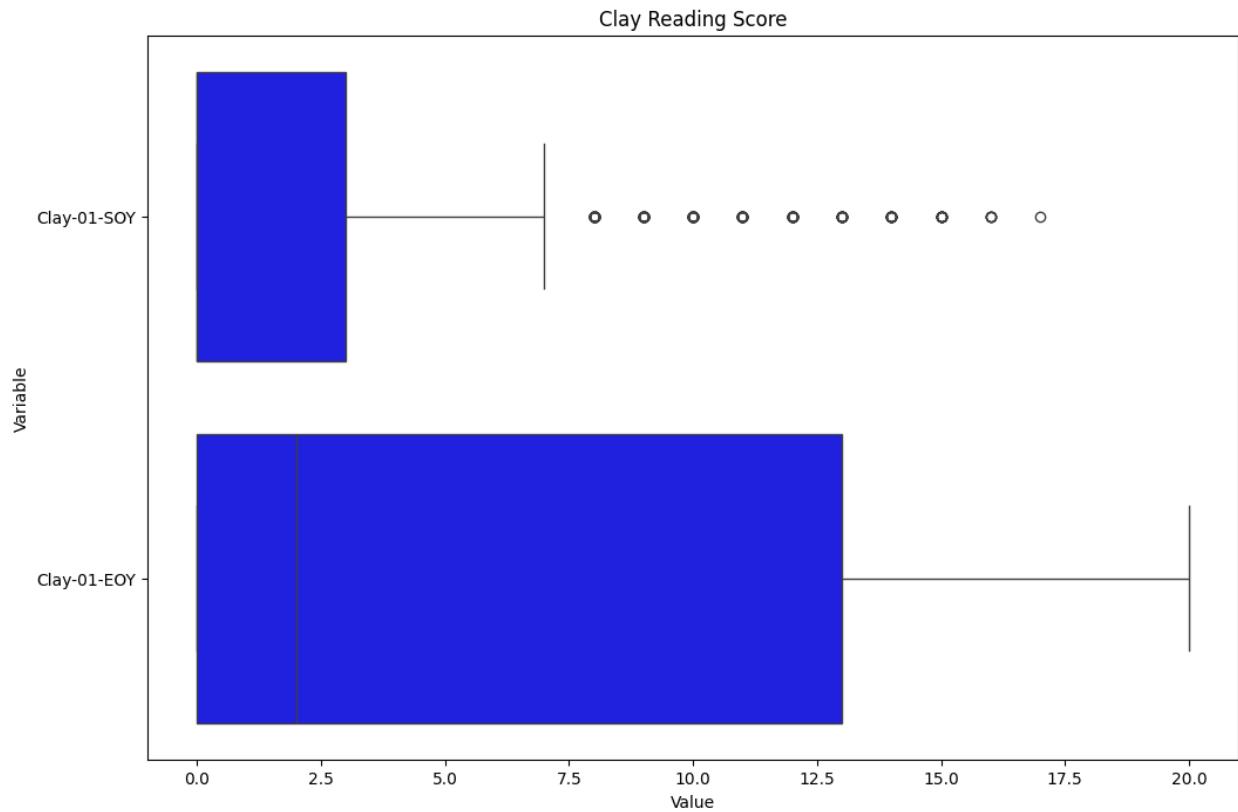


Figure 10. Box plot of Clay reading score at the start of year (SOY) and end of year (EOY) for Year 1 students

In Figure 10, the mean increased considerably from 2.65 to 6.96, and the standard deviation increased from 3.59 to 5.88, showing higher student performance variation. Clay-01-SOY data contained many outliers (244), with scores ranging from 8 to 17. However, no outliers were found in Clay-01-EOY.

Winsorisation was used to deal with Clay-01-SOY's abnormally high number of outliers. The new mean is 1.91, with a standard deviation of 2.75. Following this approach, no outliers were identified.

### g. Text Reading

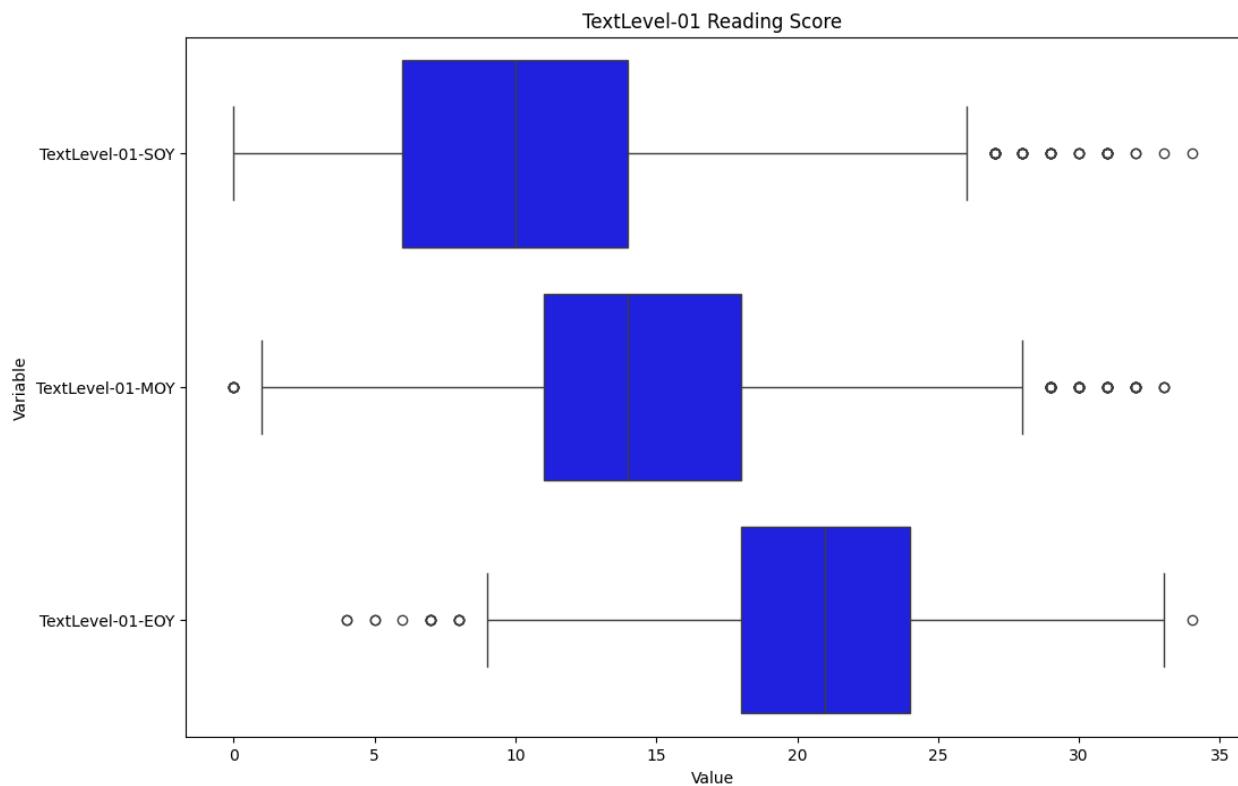


Figure 11. Box plot of Text reading score at the start of year (SOY) and end of year (EOY) for Year 1 students

From Figure 11, in Year 1, the mean 10.7 at the SOY, with a range of 34. By EOY, the average score had risen to 21.13, while the range stayed at 34, indicating significant improvement in reading proficiency. The number of outliers declined from 43 at the start to 15 by the end of the year, and the standard deviation fell from 6.1 to 4.58.

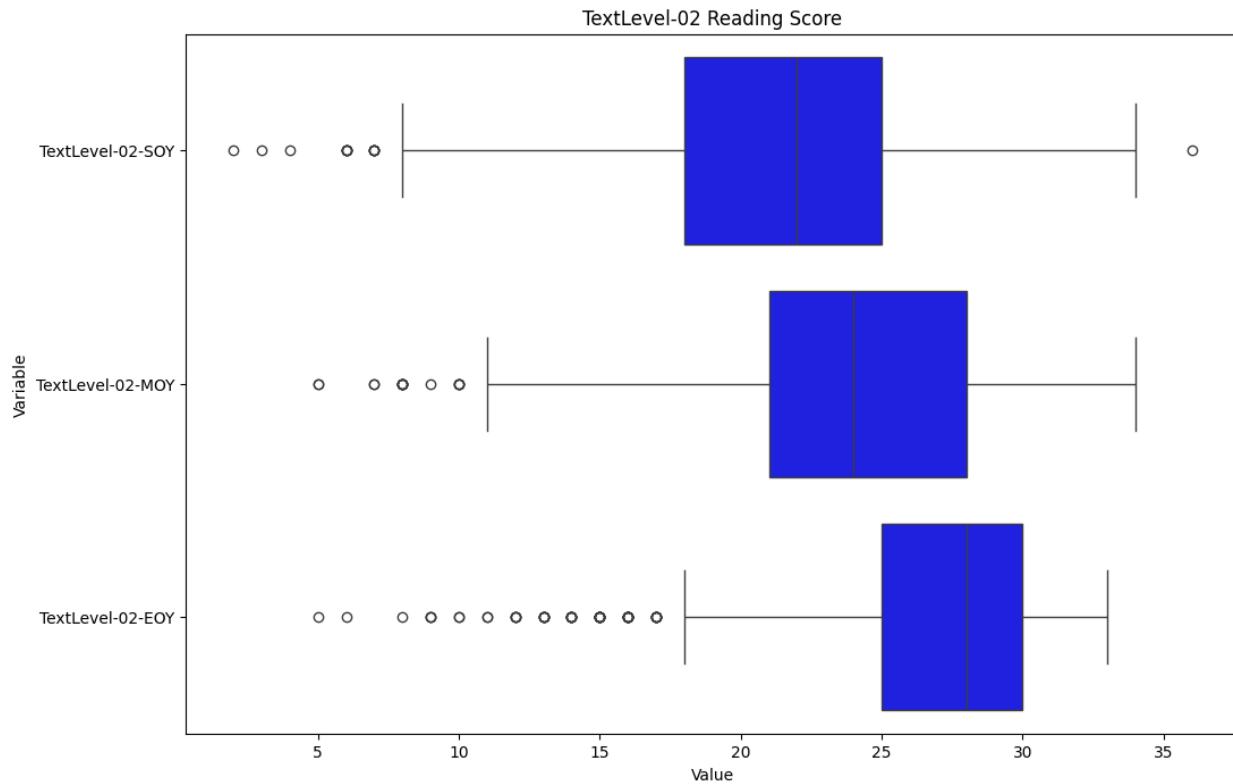
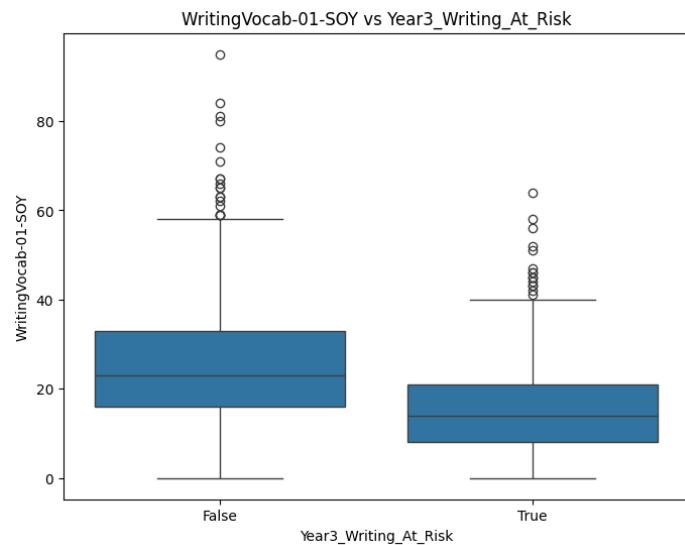


Figure 12. Box plot of Text reading score at the start of year (SOY) and end of year (EOY) for Year 2 students

From Figure 12, in Year 2, the mean at SOY was 21.79, with a range of 36. By EOY, it had risen to 27, while the range remained at 33. The number of outliers increased from 14 to 55, whereas the standard deviation fell from 5.22 to 3.77, demonstrating significant progress despite increasing score variation.

### 3. Bivariate and Multivariate Analysis

#### a. WritingVocab-01-SOY and Year3\_Writing\_At\_Risk



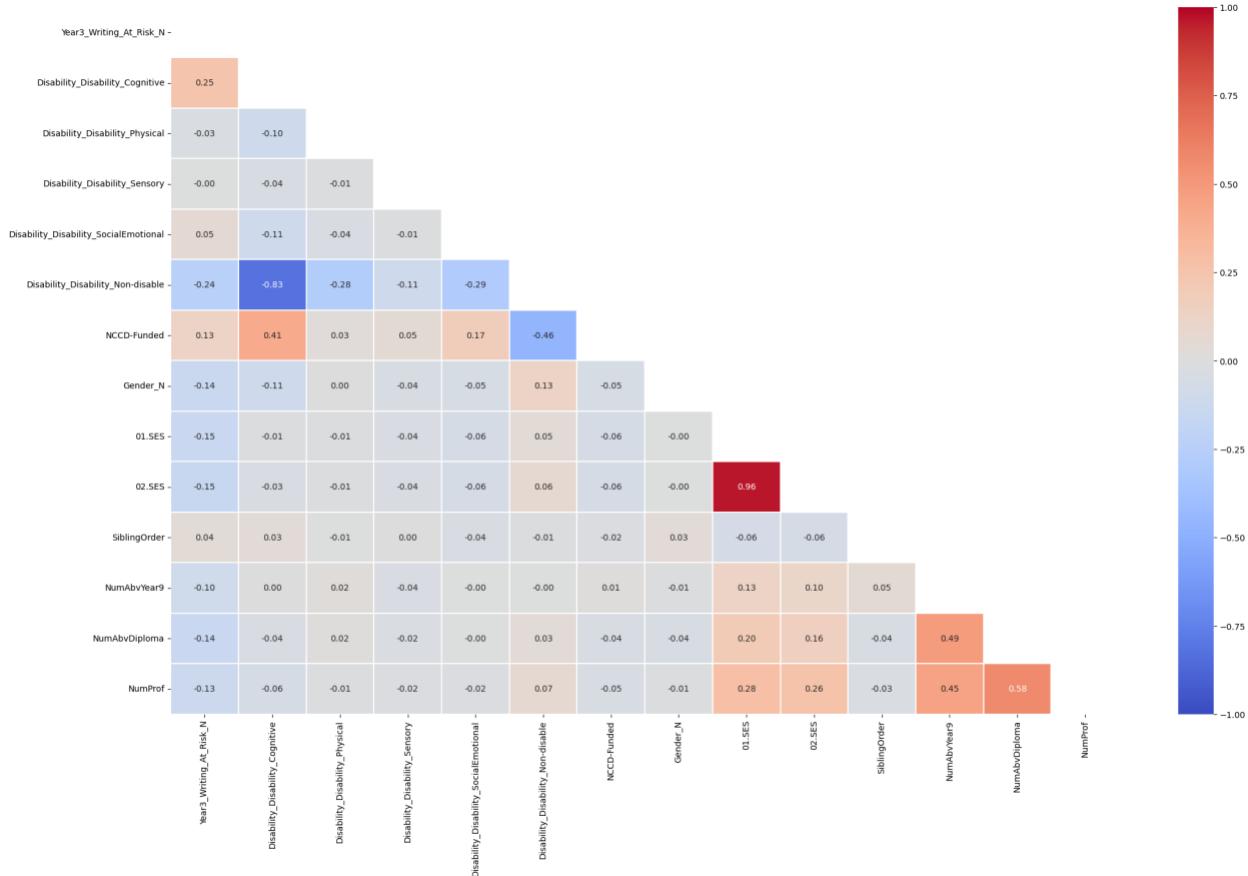
WritingVocab-01-SOY is correlated with Year3\_Writing\_At\_Risk. Figure 14 indicates that two-thirds of students not at risk of underperformance have higher average scores (25.1) than their at-risk peers (15.8). Higher Year 1 writing vocabulary correlates with lower risks in Year 3 writing, as evidenced by the negative correlation value of -0.35 – Heatmap Appendix 1.

Figure 14. Boxplots of Writing vocabulary tests against Year3\_Writing\_At\_Risk

#### b. Literacy-oriented and Numeracy-oriented vs. Year 3\_Writing\_At\_Risk

Except for the Clay reading test, Year 3\_Writing\_At\_Risk has a negative relationship with the majority of numeracy and literacy assessments (Appendix 1). Higher assessment results reduce the likelihood of underperforming in Year 3 writing. Text reading exams have the strongest correlations (-0.35 to -0.39), but numeracy assessments had weaker correlations (-0.13 to -0.21).

#### c. Disability and Year3\_Writing\_At\_Risk



**Figure 15. Heatmap between student characteristics, background, and Year3\_Writing\_At\_Risk**

According to 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 that students with cognitive disabilities are more likely to struggle with Year 3 writing. Disability\_Non-disable shows a moderate negative correlation (-0.24).

## 4. Data Conversion

Three features were converted from Categorical variables to Numerical variables for machine learning developments.

- Year3\_Writing\_At\_Risk to Year3\_Writing\_At\_Risk\_N with 0 as False and 1 as True
- Disability: Dummy-coded five categories into new variables
  - o Disability\_Disability\_Cognitive

- Disability\_Disability\_Physical
- Disability\_Disability\_Sensory
- Disability\_Disability\_SocialEmotional
- Disability\_Disability\_Non-disable
- Gender to Gender\_N with 0 Male and Female as 1

## Model development and evaluation

Two supervised machine learning models, Logistics Regression and K-Nearest Neighbour (k-means), were deployed to predict students at risk of underperforming in the Year 3 writing NAPLAN test.

Data was split into 80% for training and 20% for testing. Data was randomly selected using the stratified method and random seed of 2024 to ensure consistency and comparable outputs between the two models.

### 1. Logistic Regression

#### a. Model Justification

Year3\_At\_Writing\_At\_Risk is a categorical dichotomous variable (True/False), and logistic regression is designed to estimate the probabilities of binary outcomes.

Logistic regression is easy to interpret because one of the outputs is an equation, allowing us to assess the contribution of each predictor in influencing the probability of the outcome (at risk or not at risk).

Logistic regression is easy to train and designed for small datasets. Given the small size of our dataset, it is also easy to enhance the performance of the final model.

## b. Features Selection

Ten features were selected as predictors. They are selected because they are moderately correlated with our target features (corr value > 0.2) based on the Heatmap data in Appendix 1. These features include

- Burt-01-SOY and Burt-01-EOY
- TextLevel-01-SOY, TextLevel-01-MOY, and TextLevel-01-EOY
- TextLevel-02-SOY, TextLevel-02-MOY, and TextLevel-02-EOY
- Counting-02 and Multiplication and Division-01

Multiplication and Division-01 were selected despite a low correlation (below 0.2). The RFE method was run, and combining these nine predictors produced the best result.

Because each test has a different grading system, selected features were transformed through standard scaling. Standard scaling is more suitable in this case because it is less sensitive to outliers present in these tests.

## c. Model Results and Interpretation

$$\begin{aligned} \text{Year\_3\_Writing\_At\_Risk} = & -0.893 + -0.159 * \text{Burt-01-SOY} + -0.280 * \text{Burt-01-EOY} + -0.173 \\ & * \text{TextLevel-01-SOY} + -0.192 * \text{TextLevel-01-MOY} + -0.080 * \text{TextLevel-01-EOY} + -0.206 * \\ & \text{TextLevel-02-SOY} + -0.044 * \text{TextLevel-02-MOY} + -0.247 * \text{TextLevel-02-EOY} + -0.080 * \\ & \text{Counting-02} + -0.088 * \text{Multiplication and Division-01} \end{aligned}$$

- 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.

#### d. Model Performance

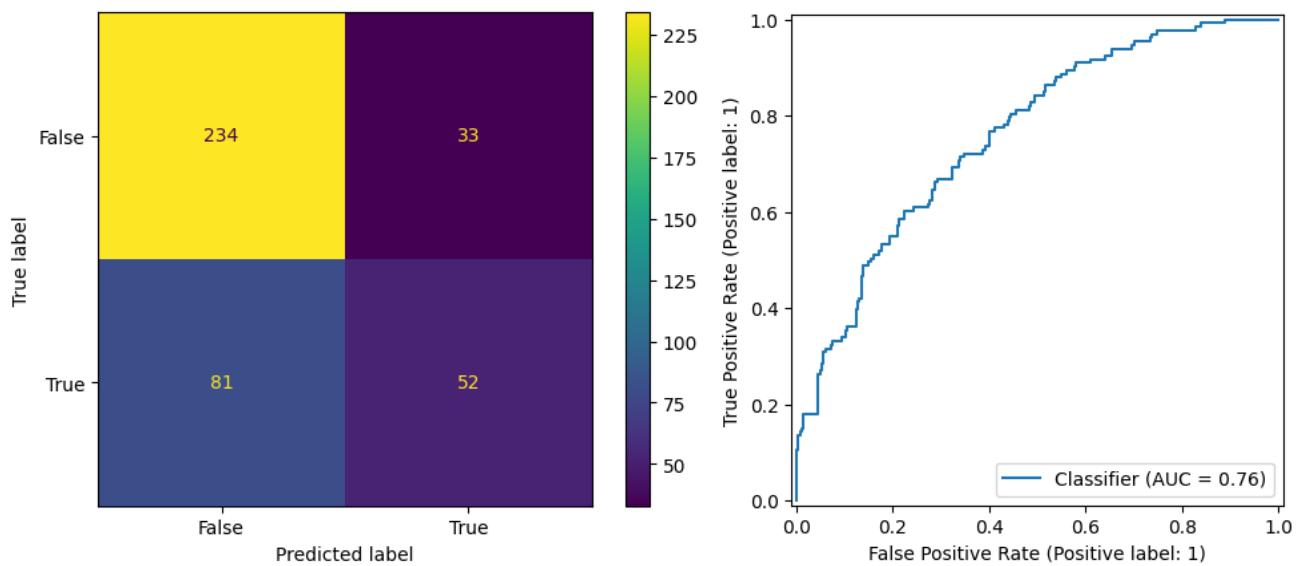


Figure 16. Confusion matrix and ROC curve (Logistic Regression)

Based on Figure 16,

- Accuracy = 71.5%: The model has moderate predictive power. It correctly classifies performance status for about 71.5%.
- Precision for “True” = 0.61. Only 61 out of 100 students are predicted by the model to be “at risk.”
- Recall for “True” = 0.39. The model can only identify 39% of the students who were actually “at risk.”

- F1 score for “True” = 0.48. This low F1 score suggests that the model struggles to consistently and accurately identify students at risk and not at risk.
- AUC = 0.76: The model predictive ability in distinguishing between at-risk and not-at-risk students is relatively good.

## 2. K-Nearest Neighbour

### a. Model Justification

K-NN is selected for model development for the following reason:

- The dataset has many features for different data types, and K-NN can handle various data types effectively. This allows us to incorporate student characteristics, such as their backgrounds, into the model, enhancing its ability to classify at-risk students.
- K-NN is easy to understand, allowing teachers and administrators to interpret and trust the result easily. The model relies on proximity to classify potential underperforming students,

### b. Features Selection

The following features were selected because they provided the best output, which was substantiated by Appendix 1 and Figure 15:

- Burt-01-SOY and Burt-01-EOY (
- TextLevel-01-SOY, TextLevel-01-MOY, and TextLevel-01-EOY
- TextLevel-02-SOY, TextLevel-02-MOY, and TextLevel-02-EOY
- Disability\_Disability\_Cognitive and Disability\_Disability\_Non-disable

Data was also transformed through standard scaling due to the different grading scales of each variable to the target feature.

### c. Model Development

The initial number of neighbours was  $k = 10$ . Since feature values were comparable thanks to standard scaling, Euclidean distance ( $p = 2$ ) was applied to measure the distance between points.

### d. Model Performance

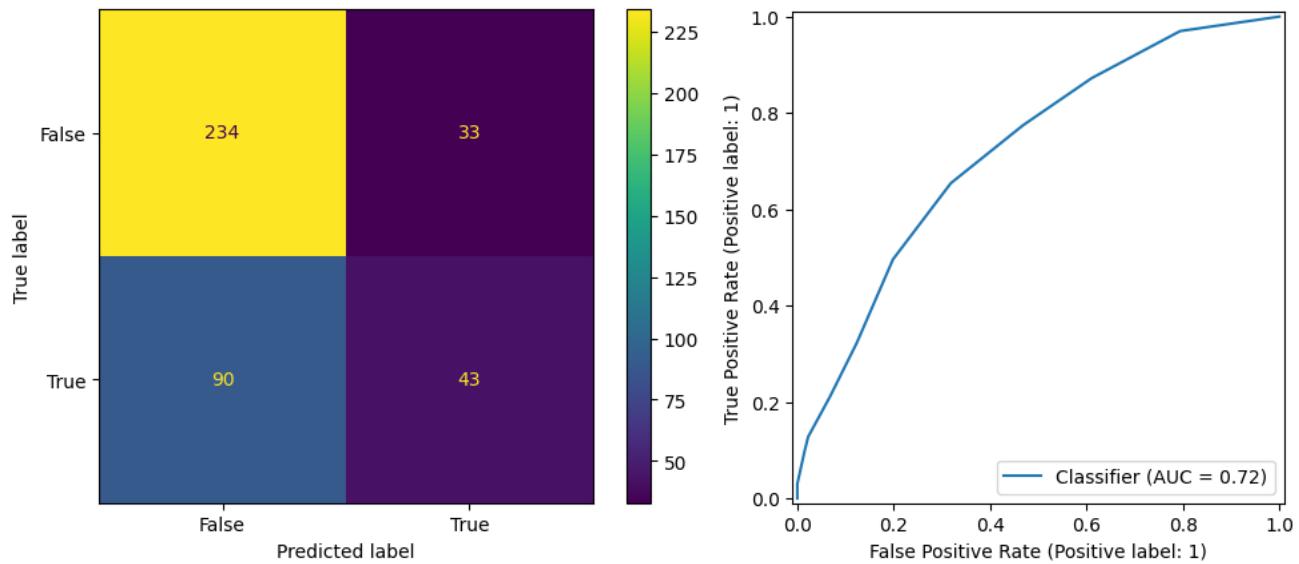


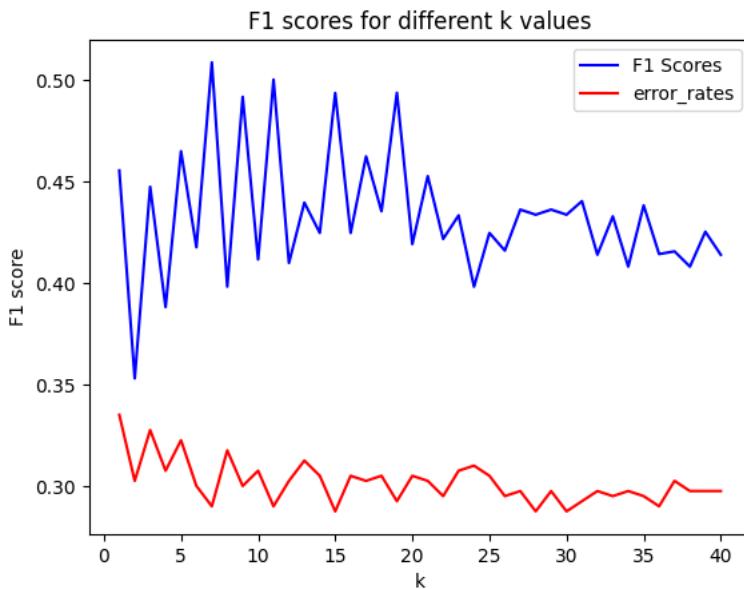
Figure 17. Confusion matrix and ROC curve (K-NN Classifier)

Based on Figure 17,

- Accuracy = 0.693: The model has moderate predictive power. It correctly classifies performance status for about 69.3%.
- Precision for “True” = 0.566. Only 57 out of 100 students are predicted by the model to be “at risk.”
- Recall for “True” = 0.323. The model can only identify 32% of the students who were actually “at risk.”
- F1 score for “True” = 0.41. This low F1 score suggests that the model struggles to consistently and accurately identify students at risk and not at risk.

- AUC = 0.72: The model's predictive ability in distinguishing between at-risk and not-at-risk students is relatively good.

### e. Optimising k



K was also optimised based on F1 score (Appendix 2), which was 7. From Figure 18, the best F1 score and accuracy were 0.508 and 0.71

Figure 18. F1 scores for different k-values

## 3. Model Comparison and Selection

Two models are compared based on their optimised and validated performance metrics.

After optimising both models by finding the best threshold, the performance metrics of each model are summarised in Table 1

	Logistic Regression	K-NN (k = 7)
Best threshold	0.29	0.29
Accuracy	0.655	0.600
Precision	0.488	0.443
Recall	0.767	0.789
F1	0.596	0.568

AUC	0.76	0.714
-----	------	-------

Table 1. Summary of Performance metrics after determining the best thresholds

From Table 1, the best threshold is 0.29. The Logistic Regression model's sensitivity to detecting at-risk students increases to 77%. The F1 score also increases substantially to 0.596, a better balance between precision and recall than the initial model. However, precision and accuracy were reduced to 49% and 66%, respectively. AUC remains at 76%.

Similarly, the best threshold for the K-NN model is 0.29. The model's sensitivity to identifying at-risk students increases to 78.9%. The F1 score also climbs significantly to 0.568, indicating a better balance of precision and recall than the first trial. However, precision and accuracy dropped to 44.3.2% and 60%, respectively. The AUC slightly reduced to 71.4%

Both models are validated using k-fold validation (k = 10):

	Logistic Regression	K-NN (k = 7)
Accuracy	0.730 +/- 0.012	0.701 +/- 0.014
F1	0.506 +/- 0.028	0.473 +/- 0.030

Table 2. Summary of performance metrics post-validation

Both models are reliable for predictive purposes with room for improvement. While K-NN generally has a slightly better recall, the logistic regression model performed better in Accuracy, Precision, F1 score, and AUC. Therefore, logistic regression is the chosen supervised machine learning model.

## 4. K-means Clustering

### a. Model Justification

K-means clustering was selected because:

- It is easy and powerful for large and small dataset
- Easy to interpret even for non-technical professional, such as school officials and teachers, and parents.

## b. Features Selection

Based on the correlation coefficient values obtained from Figure 15, the following features were selected:

- Burt-01-EOY
- TextLevel-01-SOY and TextLevel-01-EOY
- TextLevel-02-SOY and TextLevel-02-EOY
- Disability\_Disability\_Cognitive and Disability\_Disability\_Non-disable

## c. Model Development

Min-max scaling was performed because this technique ensures that no feature dominates the clustering model, which allows better groupings.

K = 4 was used to initialise the clustering model. The model was set to run ten times with different centroid initialisations with a maximum iteration of 300 and a random state of 2024.

## d. Model Performance

Within-Cluster Sum of Squares (WCSS) = 136.473. The cluster is relatively tight.

Davies Bouldin index (DBI) = 0.751 < 1, which illustrates distinct clusters. Clusters are well-separated from each other and internally compact.

Silhouette score = 0.473 < 0.5. This suggests clustering performance is relatively poor (clusters are somewhat not well-defined), and there is potential for improvement to be as close to 1 as possible, ensuring that data points inside a cluster are very similar to each other and very different from the data points in other clusters.

The first clustering model is depicted in Figure 19.

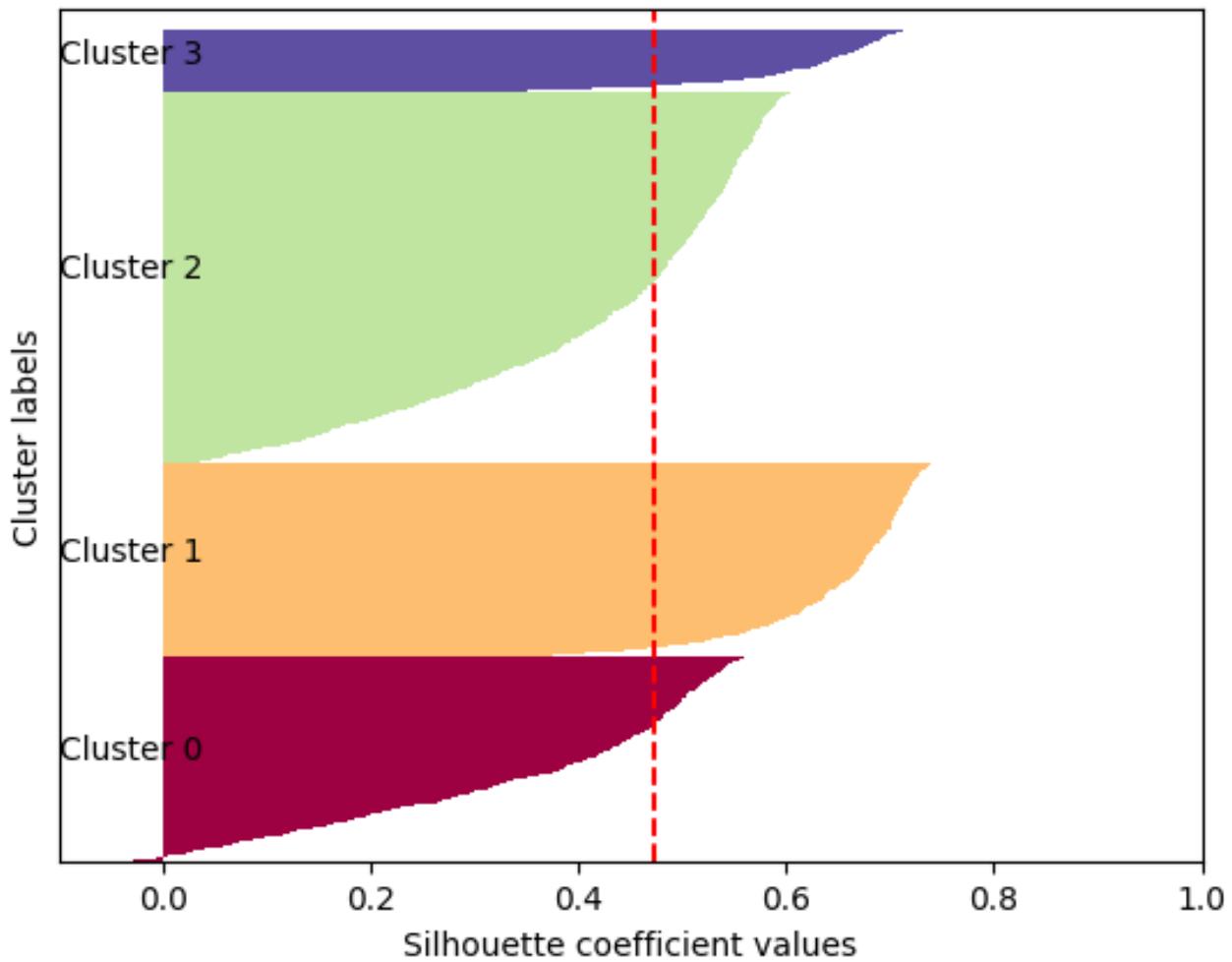


Figure 19. Silhouette diagram for four clusters

e. Model Optimisation

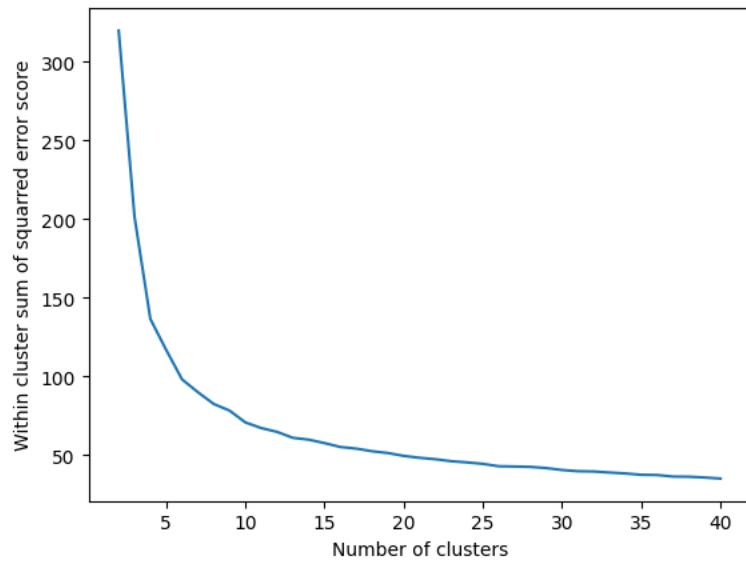


Figure 20. WCSS scores for different number of clusters

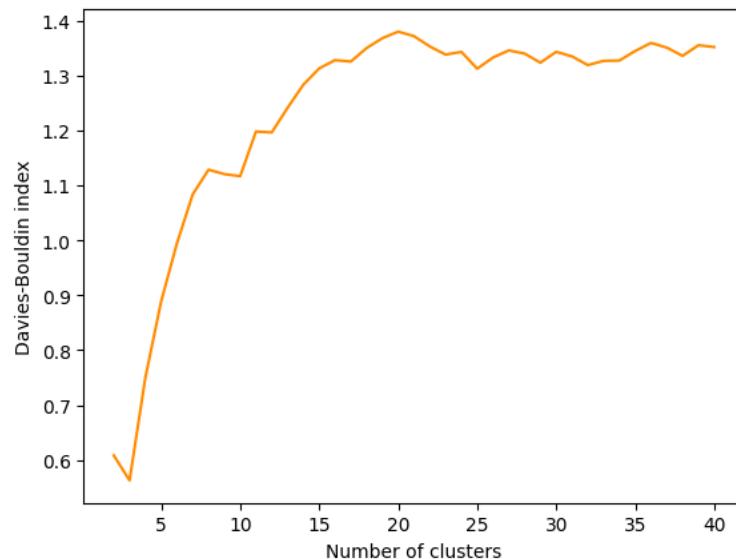
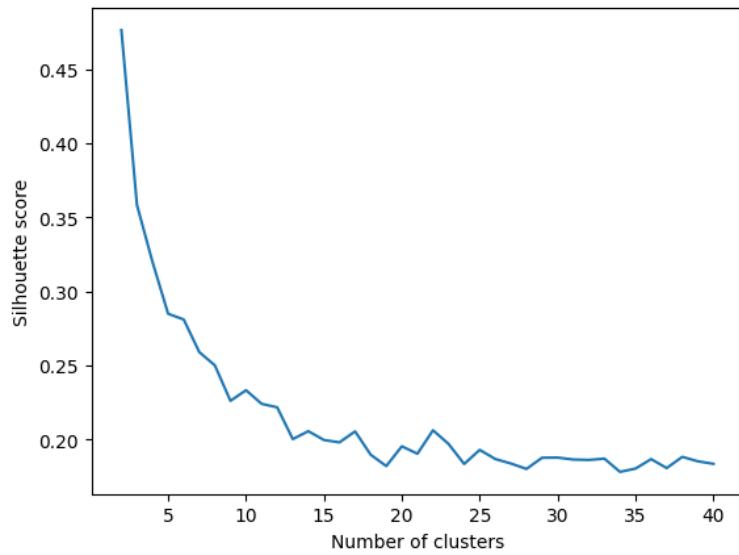


Figure 21. Davies-Bouldin index for different number of clusters

From Figure 20, the optimal k was 40 to obtain the lowest WCSS score of 34.987. However, dividing students into 40 clusters is not practical, leading to over-segmentation.

From Figure 21, the optimal k was 3 for the lowest DBI of 0.563.



From Figure 22, the best  $k$  was 2 for the optimal silhouette score of 0.652, which indicates the clustering is fair.

Figure 22. Silhouette score for different number of clusters

#### f. Model Selection

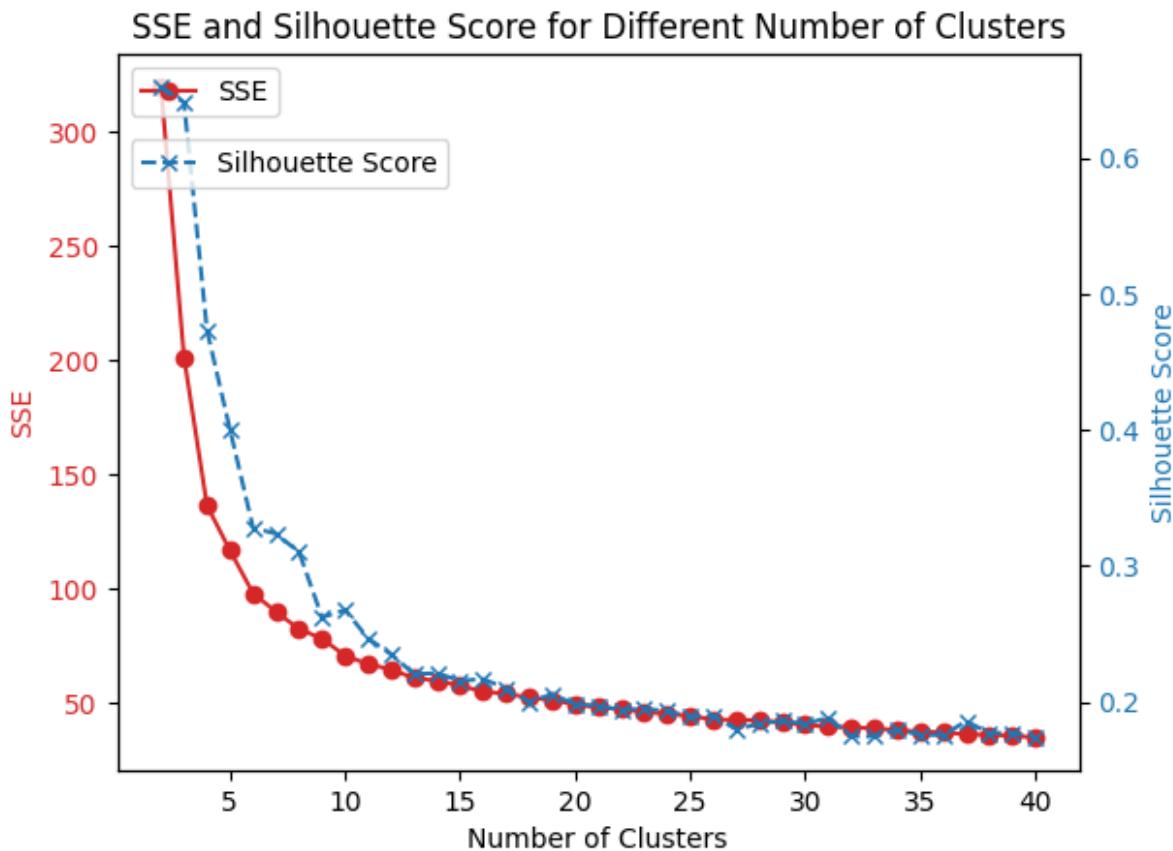


Figure 23. WCSS and Silhouette scores for different number of clusters

Figure 23 illustrates the trade-off between SSE and silhouette score in exchange for the number of clusters. K = 2 (2 clusters) was chosen for post-analysis to obtain a balance to all indices. The performance metrics are

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

The new cluster is illustrated in Figure 24

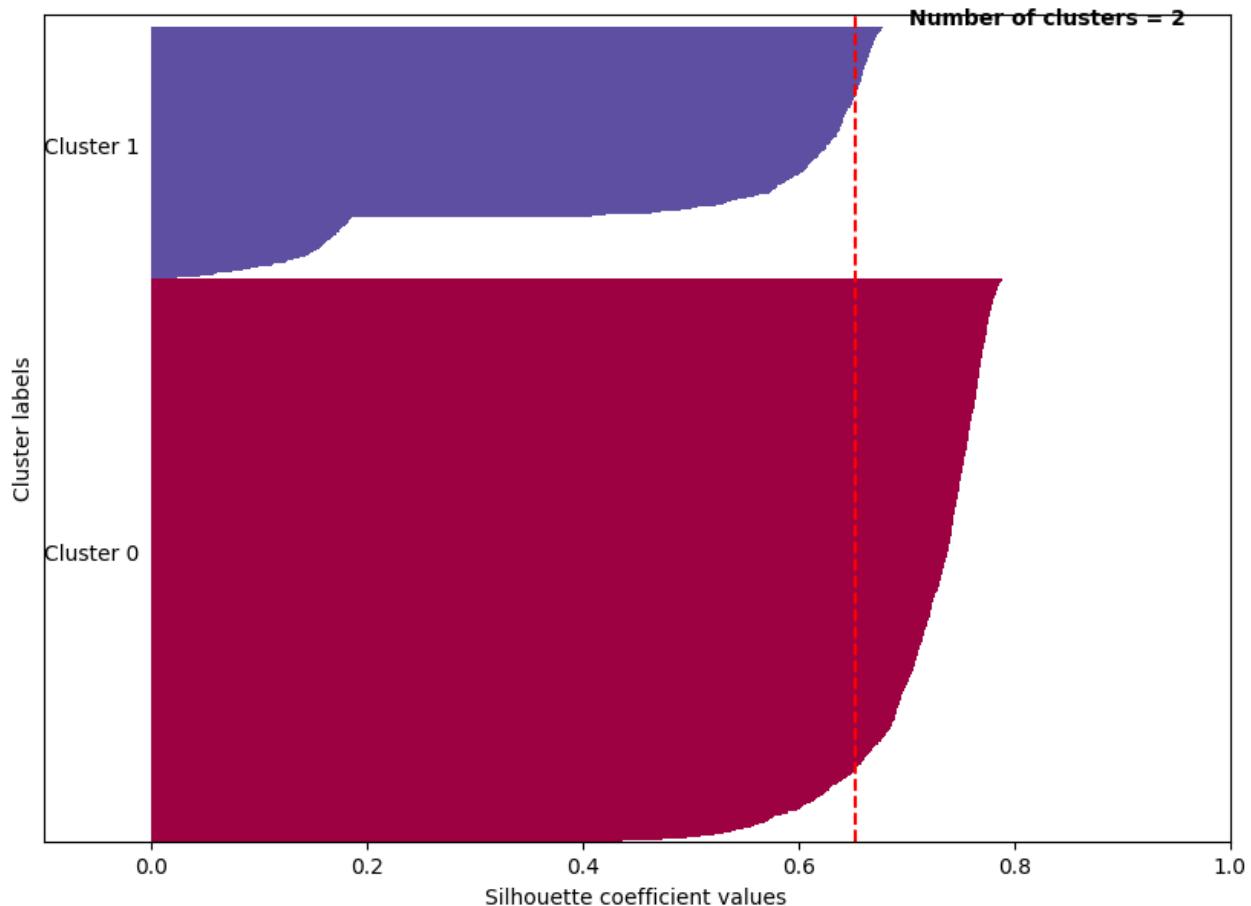


Figure 24. New cluster ( $k = 2$ )

### g. Post-Analysis

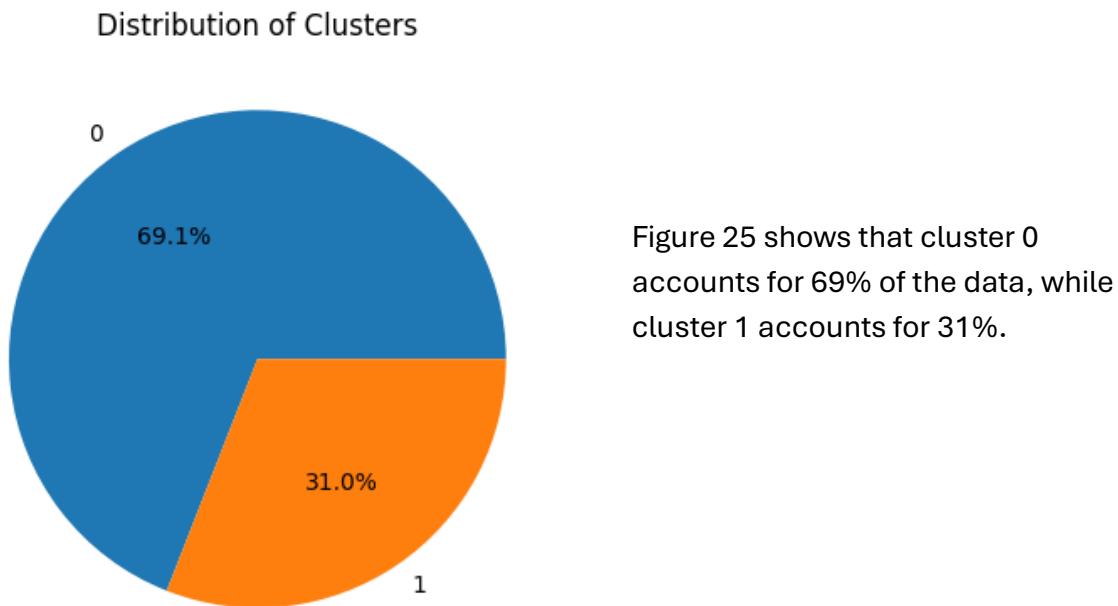


Figure 25. Distribution of clusters

Two clusters of students are profiled based on their literacy performance (Burt-01-EOY and TextLevel-02-EOY) and the presence of cognitive disabilities.

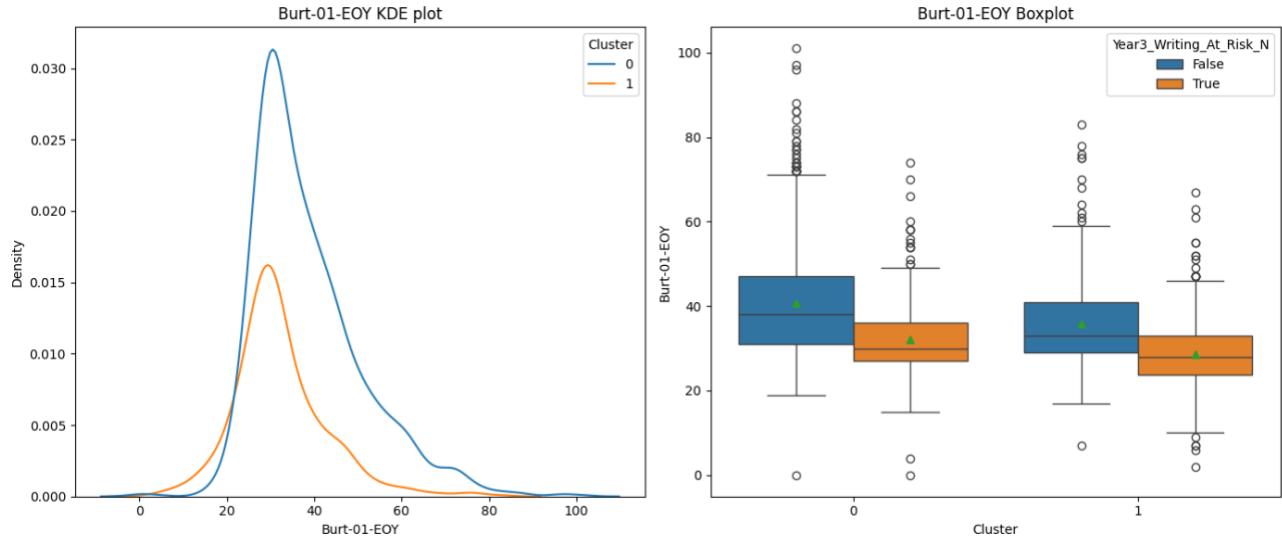


Figure 26. KDE plot and boxplot of Clusters 0 and 1 based on Burt-01-EOY

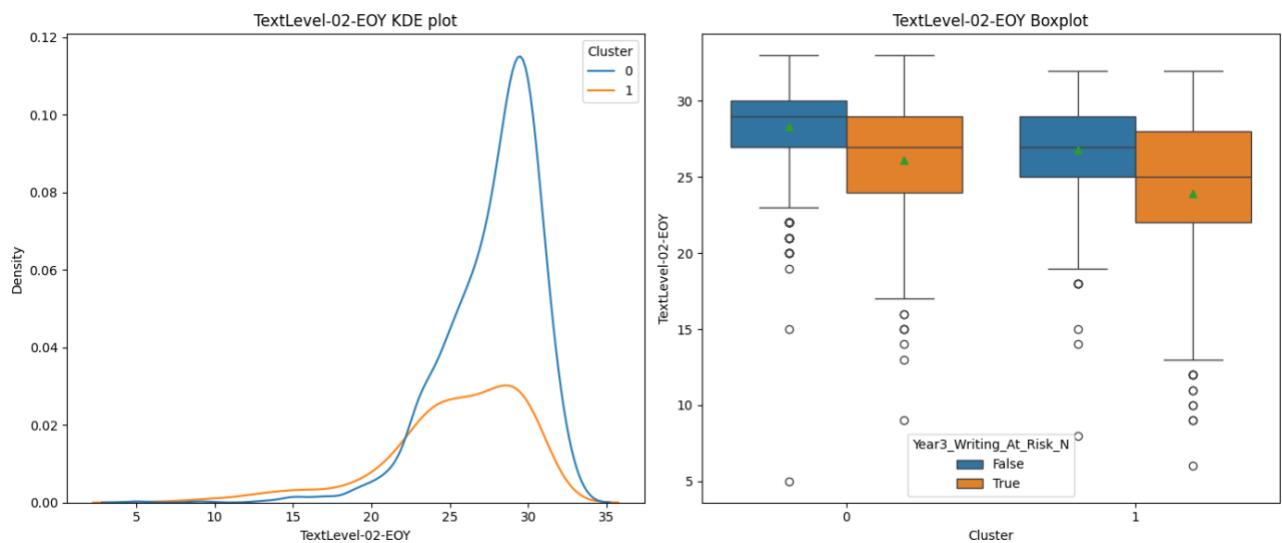


Figure 27. KDE plot and boxplot of Cluster 0 and 1 based on Text-02-EOY

Table 3 provides an overview of the two clusters.

	Cluster 0	Cluster 1
Mean Burt-01-EOY	38.5	32.2

Std Burt-01-EOY	12.4	10.6
Mean Text-02-EOY	27.7	25.4
Std Text-02-EOY	3.1	4.5
Disability status	No disabled students	Cognitively disabled students

Table 3. Characteristics of Cluster 0 and 1

## Solution recommendation

### 1. Classification Model

This report recommends using the Logistic Regression model to identify students at risk of underperformance in Year 3 writing.

The pros and cons of the model include:

Pros	Cons
Quick training time and easy to implement	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
Relatively high accuracy rate	Model is sensitive with outliers

Table 4. Pros and cons of Logistic regression model

The focus is on boosting students' Burt reading tests at the end of Year 1 and Text reading tests at the end of Year 2. Assist and instruct student who shows early signs of difficulty in these areas. Consider one-on-one or small-group reading sessions a priority. Teachers and parents will need to coordinate to track progress to ensure students learn the basic reading skills necessary for improved writing outcomes while engaging families to encourage literacy at home.

## 2. Clustering model

Based on the post-analysis of the clustering model, the report recommends dividing students into two groups.

1. Group 1: High literacy performance (no disability) with lower chance of being at risk of underperforming. Support for this group of students should focus on promoting advanced reading and writing activities such as creative writing and reading book day and competition.
2. Group 2: Low literacy performance (with cognitively disabled students) with higher chance of being at risk of underperforming. Support for this group of students should focus on
  - Technology such as visual aids, text-to-speech tools, and graphic organisers to enhance understanding
  - Communication between parents and teachers to set goals and out-of-school practice tasks
  - Pair students of Group 1 to support students in Group 2 and promote collaboration

## 3. Future engagements with the client

The data size is small and could lead to overfitting. More student data across different geographic regions and student demographics is required for better model development and more reliable results for future use.

The dataset provided does not include 2019 data. Integrating 2019 could provide more valuable insights and enhance the overall analysis.

Improve data quality

- Many features are not highly correlated with the target features, especially student backgrounds and characteristics. Data2Intel needs to collect more related data to understand the cluster better.

- A more thorough data quality check is required to address preventable mistakes such as negative score in some test scores.
- The scope of test scores needs to be specified for a more comprehensive understanding of student performance.

## Technical recommendations

### 1. Programming Language

The programming language is Python, and the computing environment is Google Collab. Google Collab was chosen because it is highly accessible to all stakeholders. No setup and hardware are required. Python is free, powerful, and flexible, with lots of libraries to assist the data analysis process.

The following software libraries were used:

- Data manipulation and analysis: Pandas and Numpy
- Data visualisation: Matplotlib, Seaborn, Matplotlib ticker, Matplotlib CM
- Machine learning and modelling: Scikit-learn for splitting and scaling data, and for the two supervised and one unsupervised machine learning models.
- Model evaluation metrics: sklearn import metrics

### 2. Machine Learning Diagram

#### a. Data Collection

Secondary data: Data from 40 Australian schools is collected and aggregated by Data2Intel.

#### b. Data Pre-processing and Cleaning

Inspect missing data and data types

Handling errors, missing data and features with an exceptionally high number of outliers.

Scale data via standard scaling (for classification model) and min-max scaling (for clustering model) and convert categorical variables into numerical variables for model development

### c. Exploratory Data Analysis

Explore student distribution based on their test scores from literacy and numeracy-oriented assessments.

Explore student at-risk and not-at-risk writing trends and their distribution in Year 3.

Explore student characteristics and background

Explore tests relationship and some student characteristics with Year3\_Writing\_At\_Risk

### d. Model Training and Evaluation

Two supervised machine learning models—Logistic Regression and K-NN—predict and classify students at risk of writing, and one clustering model using k-means identifies student groups based on their literacy and disability profiles.

Evaluate classification models using confusion matrix, ROC curve, accuracy, precision, recall, and F1 scores.

Evaluate clustering models via WCSS score, DBI index, and silhouette score.

### e. Model Deployment

Deploy the chosen logistic regression model to classify students at full data scale from 40 schools.

Deploy the chosen clustering profile to group students for target intervention based on their Burt and Text reading test scores.

### f. Model Monitor and Feedback

Feedback loops will be established with teachers and school officials to help adjust predictions based on real-world results. As new student data becomes available, the

models will be retrained to maintain their accuracy and applicability over time. The analyst team and Data2Intel will also monitor changes in data distribution to detect anomalies and guarantee model reliability.

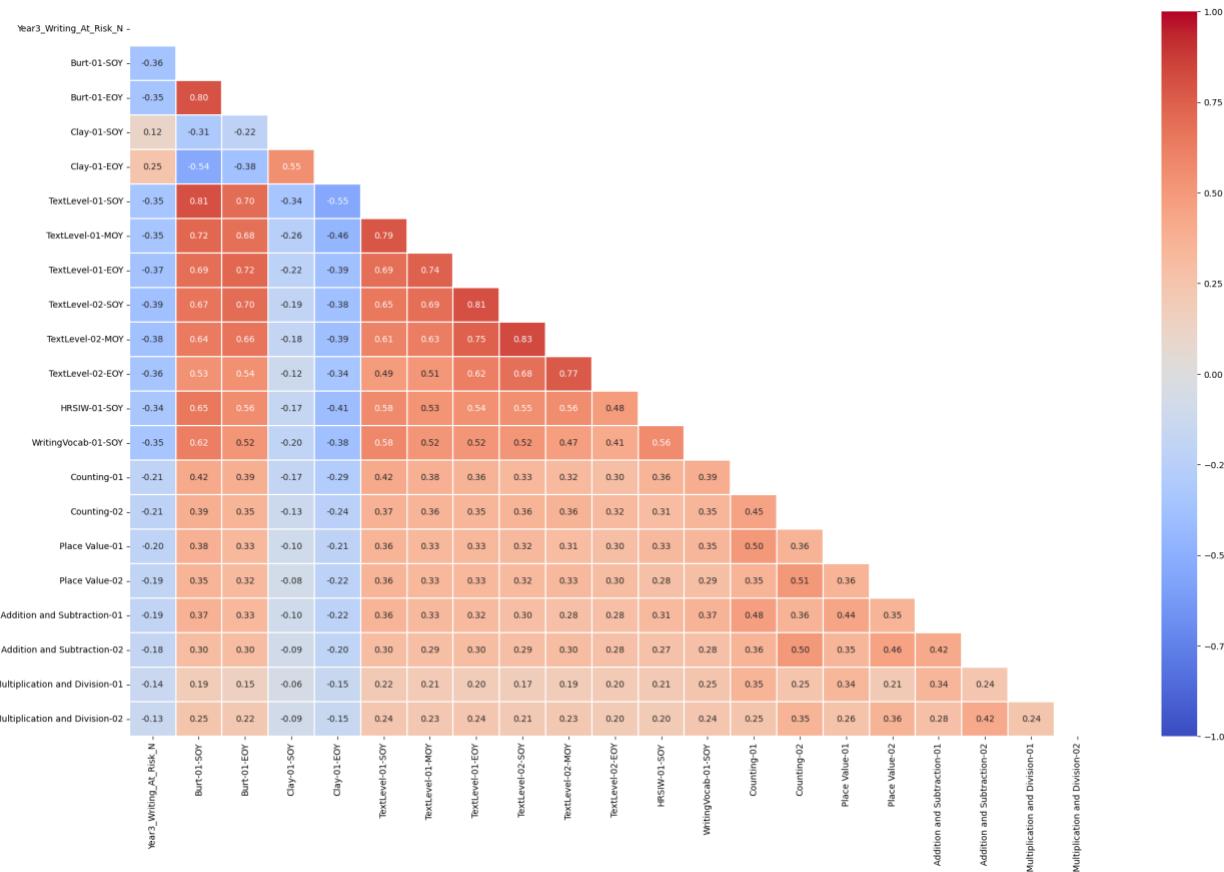
### 3. Maintaining Model's Accuracy and Relevance

To maintain the model's accuracy and relevance over time, the following steps must be fulfilled.

- Update data and automate the process for data collection
- Data quality checks must be conducted regularly to detect shifts or missing values
- Expand the dataset and include more diverse student populations to improve the model generalisability.
- Feedback loop and multiway-way communication between Data2Intel, business analysts, educators and students to ensure the model align with real-world results.
- Consider excluding unrelated features with low correlation values with the target features such as NumAbvYear 9 and SibingOrder

## Appendix

### Appendix 1. Heatmap between literacy-oriented and numeracy-oriented assessments with Year3\_Writing\_At\_Risk



### Appendix 2. The decision to prioritise recall over precision

The report prioritised "recall" to identify as many at-risk students as possible for early intervention. Higher recall rates mean fewer false negatives, allowing for identifying most students who need support and assistance from teachers to improve their writing and reading skills. False negatives, or missing at-risk students, can have detrimental long-term effects on students' academic success, increasing learning disparities and possibly

influencing future studies. Although a higher precision rate would lower false positives, in this case, it may not be as important because misclassifying a student as at risk is less costly than neglecting to promptly assist a student who needs support from a teacher. As a result, the best threshold is determined based on F1 score for both Logistic Regression and K-NN models.