|  |
| --- |
|  |

|  |
| --- |
| Machine Learning in Business  MIS710 – A1  MIS710 – A2 |

|  |  |
| --- | --- |
| Sajana Nisal Wijethilaka Thalagodage, 224748948 Email address |  |



[SECTION 2](#_Toc1855681488)

[Executive Summary 3](#_Toc409078313)

[Introduction 3](#_Toc480207072)

[Business Understanding (BACCM Framework) 3](#_Toc2069465379)

[Insights from Exploratory Data Analysis (EDA) 4](#_Toc253300466)

[SES Backgrounds in Year 1 and Year 2 4](#_Toc1842710805)

[Reading Skills in Year 1 and Year 2 4](#_Toc1416919207)

[Writing Skills in Year 1 4](#_Toc167121273)

[Relationship Between Literacy and Numeracy Skills 5](#_Toc2015487253)

[Disability Conditions and Year 3 Writing At Risk 5](#_Toc1153417249)

[Additional Insights 5](#_Toc52386149)

[Proposed Machine Learning Solution 5](#_Toc489527135)

[Selected Machine Learning Model 6](#_Toc1617393299)

[Interpretation of Model Performance 6](#_Toc1761515921)

[Recommendations and Conclusions 6](#_Toc425204838)

[Recommendations of Business Applications 6](#_Toc477458769)

[Potential Benefits to Stakeholders 7](#_Toc1559488771)

[Changes to Business Processes and Impacts 7](#_Toc1714021600)

[Recommendations for Further Improvements 7](#_Toc240536410)

[Conclusion 7](#_Toc1338802129)

[SECTION 8](#_Toc1785056346)

|  |
| --- |
| SECTION |

Executive Summary

This report provides insights into student performance based on the LA4PSchools.csv dataset, using machine learning to predict which students are at risk of underperforming in Year 3 writing assessments. The analysis answers key business questions about students’ SES backgrounds, literacy and numeracy skills, disability conditions, and their relationship to future writing performance. The Random Forest model was identified as the most effective tool for predicting at-risk students, achieving an accuracy of 82%. Recommendations include deploying this model to improve early interventions, enhance resource allocation, and drive personalized learning strategies.

Introduction

Business Understanding (BACCM Framework)

The Business Analysis Core Concept Model (BACCM) is a framework that helps structure business understanding by considering six core concepts: stakeholders, needs, solutions, value, context, and change.

1. Stakeholders:

The key stakeholders are primary school administrators, teachers, and educational policymakers. These stakeholders are responsible for identifying students at risk of underperforming in writing and implementing interventions to improve academic outcomes.

1. Needs:

The primary need is a system that can accurately identify students at risk of underperforming in Year 3 writing assessments. Early identification allows schools to intervene with targeted learning plans to enhance students' literacy skills before their performance declines significantly.

1. Solutions:

The machine learning solution developed in this project provides a predictive model that uses student data from Years 1 and 2 to assess the likelihood of underperformance in Year 3 writing. The model empowers educators to allocate resources effectively and implement personalized learning strategies.

1. Value:

The value proposition lies in the ability to identify at-risk students early, which can lead to improved academic outcomes, increased literacy rates, and more efficient use of teaching resources. The system reduces the risk of long-term academic failure, fostering a supportive educational environment.

1. Context:

The project takes place within the context of primary school education, specifically targeting literacy improvement. Schools face constraints related to resource availability, requiring data-driven methods to prioritize students in need of support.

1. Change:

Implementing the machine learning model will lead to changes in how schools identify and support at-risk students. Schools will transition from reactive intervention methods to a proactive, data-driven approach that optimizes learning outcomes.

Insights from Exploratory Data Analysis (EDA)

SES Backgrounds in Year 1 and Year 2

The dataset includes information on the socioeconomic status (SES) of students in both Year 1 and Year 2. The national average SES score for Catholic schools is 100, and for independent schools, it is 102, according to the Australian Department of Education and Training. The SES scores in the dataset show that most students fall within this range, but there are notable outliers with lower SES scores, indicating that some students may come from more disadvantaged backgrounds.

Reading Skills in Year 1 and Year 2

Students' reading skills were assessed using Burt Reading Scores at the start (SOY) and end (EOY) of both Year 1 and Year 2. The analysis revealed that while most students demonstrated improvement in reading skills from Year 1 to Year 2, a significant portion of students started with below-average scores. These students showed minimal improvement, particularly those who were identified as at risk in the Year3\_Writing\_At\_Risk variable.

Writing Skills in Year 1

The variable WritingVocab-01-SOY reflects students’ writing skills at the start of Year 1. Initial analysis showed a positive correlation between WritingVocab-01-SOY and Year 3 writing performance, with students who scored poorly in Year 1 more likely to be classified as at risk in Year 3. This finding underscores the importance of early intervention in writing skills development, as early deficits often carry forward.

Relationship Between Literacy and Numeracy Skills

A correlation analysis was performed to assess the relationship between students' literacy and numeracy skills. The analysis indicated a moderate correlation between literacy and numeracy, meaning students who perform well in literacy often do well in numeracy, and vice versa. Both literacy and numeracy skills were also found to be related to the Year3\_Writing\_At\_Risk variable, further highlighting the interconnectedness of core academic skills in student performance.

Disability Conditions and Year 3 Writing At Risk

The dataset includes information about students' disability conditions. A significant finding from the analysis is that students with disabilities were more likely to be categorized as at risk of underperforming in Year 3 writing assessments. This highlights the need for additional support structures for students with disabilities to ensure they have equal opportunities to succeed academically.

Additional Insights

Additional insights from the clustering analysis suggest that students can be segmented into two distinct clusters based on their literacy and numeracy scores. One cluster represents high-performing students, while the other represents students who are likely to underperform without targeted interventions. These insights emphasize the value of differentiated teaching approaches, where schools can focus efforts on the students most in need.

Proposed Machine Learning Solution

Selected Machine Learning Model

After evaluating several machine learning models, the Random Forest Classifier was selected as the best-performing model. This model achieved an accuracy of 82% and demonstrated balanced precision and recall metrics, ensuring that both at-risk and not-at-risk students were accurately identified.

Interpretation of Model Performance

The Random Forest model's ability to handle complex relationships between features, such as literacy scores, SES background, and disability status, made it an ideal choice. Its performance metrics show that it can accurately identify students who need support while minimizing false positives. The model’s interpretability also allows educators to understand the key features driving predictions, enabling informed decision-making.

Pros:

* High accuracy in predicting at-risk students.
* Handles complex feature interactions (e.g., literacy, SES, and disabilities).
* Robust to overfitting due to ensemble methods.

Cons:

* Requires regular retraining to maintain accuracy as new student data becomes available.
* May need further optimization to account for any potential biases related to demographic factors.

Recommendations and Conclusions

Recommendations of Business Applications

The Random Forest predictive model should be integrated into the school’s data systems to allow real-time predictions based on updated student data. This would enable schools to:

* Proactively identify at-risk students: Teachers can intervene early with personalized learning strategies tailored to students’ needs.
* Target resources efficiently: Schools can allocate teaching assistants, tutoring sessions, and other resources to students most in need.

Potential Benefits to Stakeholders

* For students: Early identification of writing difficulties allows for more focused and effective interventions, improving literacy rates and overall academic success.
* For teachers: Teachers can leverage the model’s predictions to customize teaching plans, ensuring that no student is left behind.
* For administrators: The model aids in data-driven decision-making, ensuring resources are used efficiently to maximize educational outcomes.

Changes to Business Processes and Impacts

Implementing the predictive model would result in several changes to the current educational processes:

* Data-driven decision-making: Schools would shift from reactive identification of struggling students to a proactive approach, allowing for earlier interventions.
* Changes in resource allocation: Teachers and support staff would focus more intensively on at-risk students, as identified by the model.
* Enhanced focus on literacy skills: Schools may implement additional literacy programs, informed by the model’s predictions.

Recommendations for Further Improvements

* Model retraining: To maintain accuracy, the model should be retrained every six months with updated student data.
* Bias monitoring: Regular audits should be conducted to ensure the model does not introduce biases, particularly regarding SES background and disability conditions.
* Incorporate additional features: Future models could include new features, such as parental involvement or extracurricular activities, to further improve predictive accuracy.

Conclusion

The deployment of the Random Forest predictive model offers significant potential to improve student outcomes by enabling schools to identify and support at-risk students early. By adopting this data-driven approach, educational institutions can make more informed decisions, optimize resource allocation, and foster an inclusive learning environment that supports all students, particularly those most in need of assistance.

|  |
| --- |
| SECTION |