Project Overview

You’re engaged to work with Data2Intel, an Australian learning analytics consulting service that specialises in delivering various data analytics services within the primary education sector. The current project aims to predict primary school students who are at risk of underperforming in numeracy in Year 3. This initiative is part of a broader effort by a consortium of forty primary schools to enhance educational outcomes and provide targeted support to students in need.

# Context/Scenario

Australian school education comprises K-12, a system akin to those in many other countries. Primary schools encompass scholastic years K-6, while secondary schools cover years 7-12. In Australia, most Kindergarten students must be at least 5 years old by January of the calendar year they commence schooling. The early years of schooling, typically K-2 (Kindergarten to Year 2), encompass children aged 5-7 years. The National Assessment Program – Literacy and Numeracy (NAPLAN) tests are annual assessments for students in Years 3, 5, 7, and 9 to provide a key benchmark for assessing students' foundational literacy and numeracy skills. Research, although more limited in scope compared to secondary and tertiary education, underscores the profound importance and influence of these early years on future academic performance, employability, wellbeing, and career progression. The foundational skills acquired during these formative years set the stage for long-term educational and professional success. You are provided with a dataset of 2,000 students across over forty schools. The dataset focuses on their reading and numeracy skills during the early years of Year 1 and Year 2. These skills were measured through localised, formative assessments, which, while validated and consistent, are not solely "pen and paper" tests.

Given the young age of the students, these assessments often include dialogue or interview-based evaluations administered by trained teachers. In the dataset, students at risk of underperforming in numeracy, Year3\_Numeracy\_At\_Risk, were determined by their NAPLAN results in Year 3. In addition, the dataset also provides students’ demographic and family backgrounds as well as disability conditions. Further details can be found in the supplied data description. This comprehensive dataset spans five continuous years, from 2016 to 2020, and has been curated for learning support and research purposes. Importantly, there is no missing data, as records with incomplete information have been removed for this exercise. However, you are still required to check data quality and preprocess the data as needed.

# Goal of Project

You are helping them **predict which students in Year 3 (around 8–9 years old)** might **struggle in numeracy (math)** in the future.

So that **teachers and schools can give extra help early on** — before it’s too late.

# Specific Requirements

You are tasked with performing two analytical tasks: 1/ uncovering data insights and 2/ exploring machine learning opportunities, and reporting the findings to Data2Intel.

Regarding the first task - uncovering data insights, you are required to respond to the following enquiries:

* Student demographics analysis: Analyse student demographics factors to gain insights about the students included in the dataset and report findings and insights.
* What are students’ numeracy skills in Year 1, including Counting-01, Place Value-01, Addition and Subtraction-01, and Multiplication and Division-01. Is there a relationship between students’ numeracy skills in Year 1 and Year3\_Numeracy\_At\_Risk?
* What are students’ reading skills, for example TextLevel and Writing Vocabulary in Year 1? Are there relationships between these and their Year3\_Numeracy\_At\_Risk?
* Are students’ literacy skills and numeracy skills related, in Year 1 and in Year 2? Are there relationships between these and their Year3\_Numeracy\_At\_Risk?
* Describe the students’ disability conditions in the dataset. Are there relationships between these conditions and their Year3\_Numeracy\_At\_Risk?
* Are there other insights that might inform early interventions to improve students’ numeracy skills?

Regarding the second task - machine learning opportunities, you are required to

* . Develop and evaluate two predictive models to identify students at risk of underperforming in numeracy in Year 3 (Year3\_ Numeracy\_At\_Risk).
* Develop one clustering model to explore possible clusters of students.

Based on your findings from the above both tasks, you are required to provide actionable insights and recommendations to primary schools and educators to implement targeted interventions and support mechanisms. In addition, you should advise the Client of the potential ethical and legal implications of the models.

**Dataset provided:** LA4Schools.csv

**Data description provided:** LA4Schools data description

The dataset has been prepared and provided by Data2Intel, mirrors the real-life data, and has been preprocessed by Associate Professor Lemai Nguyen for assessment purposes.

# Deliverable Requirements

You are required to:

* Develop your business and data understandings using BACCM.
* Prepare and explore the provided dataset, cleanse and pre-process data as needed. Undertake an exploratory data analysis (EDA) to respond to the client’s six questions.
* Undertake supervised machine learning model development, evaluation, comparison and selection. Two predictive models should be developed, tested, and compared.
* Undertake unsupervised machine learning using clustering analytics.
* Develop two reports:
  + The first technical report (Part A) should present your EDA (Exploratory Data Analysis) and machine learning findings to Peter Vo.
  + The second consultancy report (Part B), for Elaine Race, Director of Education and Engagement, should present responses to the six (6) specific requests about data, insight from clustering analytics, and a predictive machine learning model.

Format and present your report professionally. Two sample report templates are provided under Assessment Resources.