# FIT1043 Assignment 2: Specification

Due date: Wednesday 2nd October 2024- 11:55 pm

# Aim

The main aim of this assignment is to conduct predictive analytics, by building predictive models on a dataset using Python in the Jupyter Notebook environment.

This assignment will test your ability to:

1. Read and describe the data using basic statistics,
2. Split the dataset into training and testing,
3. Conduct multi-class classification using Support Vector Machine (SVM)**,
4. Evaluate and compare predictive models,
5. Explore different datasets and select a particular dataset that meets certain criteria
6. Deal with missing data,
7. Conduct clustering using k-means and
8. Communicate your findings in your report/video recording.

** Not taught in this unit, you are to explore and elaborate these in your report submission. This will serve as a gentle introduction to lifelong learning, encouraging you to learn independently.

# Data

We will explore the following datasets in **Task A** (plus a dataset of your choice in **Task B**):

1. Student_List_A2.csv
2. Student_List_A2_Submission.csv

**Format:** each file is a comma separated (CSV) file
**Description:** These two datasets contain information on high school students, detailing their study habits, parental involvement, and academic performance. They are a modified version of a dataset shared by Rabie El Kharoua[1]. These two files are available on the unit Moodle site under Assessments.

# Hand-in Requirements

Please hand in a **PDF file** containing your code, answers and explanations to questions, a **Jupyter notebook file (.ipynb)** containing your Python code to all the questions, a **CSV file** for your predictions in Task A5 and a **video file**:

- The PDF file should contain answers and explanations to the questions.
    - You can use Microsoft Word or other word processing software to format your submission. Alternatively, generate your PDF from your jupyter notebook formatted using markdown. Either way save the final copy to a PDF before submitting.
    - Make sure to include screenshots/images of the graphs you generate. Also, do NOT include screenshots of your code if using Microsoft Word or other word processing software.

- The .ipynb file should contain:
    - **A copy of your work using python code** to answer all the questions.

[1] DOI: 10.34740/kaggle/ds/5195702

- The .csv file should contain:
    - your predictions in Task A5.
- The video file should contain:
    - An **up to 3-minute recording** of yourself explaining your answers to Task B1. You can use Zoom to prepare your recording. Please see Task B for more details.

You will need to submit **four separate** files (i.e., .pdf file, .ipynb file, .csv file and your video file). Zip, rar or any other similar file compression format **is not acceptable** and will have a **penalty of 10%**.

# Assignment Tasks:

Note: You need to use Python to complete all tasks.

## Task A: Data Wrangling and Analysis

### A1. Data Wrangling (4 marks)

1. Read the '**Student_List_A2.csv**' file and list the column names.
2. In this dataset, 'GradeClass' column contains the classification of students' grades based on GPA, where:

| Numerical grade classification in GradeClass | Letter grade | Meaning |
|---|---|---|
| 0 | 'A' | GPA >=3.5 |
| 1 | 'B' | 3.0 <= GPA < 3.5 |
| 2 | 'C' | 2.5 <= GPA < 3.0 |
| 3 | 'D' | 2.0 <= GPA < 2.5 |
| 4 | 'F' | GPA < 2.0 |

   Replace the numerical grade classifications (0, 1, 2, 3, 4) in the 'GradeClass' column with their corresponding letter grades ('A', 'B', 'C', 'D', 'F').

3. Can you identify any missing values in the columns of this dataset? If so, replace the missing values with the median value of the relevant column where you find missing values.
4. Identify a data quality problem related to the 'Absences' column and delete the rows that exhibit this problem. Refer to Week 4 for information on data quality problems.
5. Examine the 'GPA' and 'GradeClass' columns together for additional data quality issues. Propose an appropriate solution for these issues and resolve them.

**A2. Supervised Learning (1.5 marks)**

1. Explain supervised machine learning, the notion of labelled data, and train and test datasets.
2. Use the wrangled data from A1 and separate the features and the label. Note that:
    - the label, in this case, is the 'GradeClass'
    - studentID is not logically a useful predictor of a student's grade so should not be used as a feature
    - GPA is translated to GradeClass. They both represent the same thing so GPA should not be used as a feature.
    - Use the rest of the features as predictors.
3. Use the sklearn.model_selection.train_test_split function to split your data for training and testing (Keep 80% of the data for training).

**A3. Classification (training) (3 marks)**

1. In preparation for classification, your data should be normalised/scaled.
    a. Describe what you understand from this need to normalise data (this is in your Week 7 applied session).
    b. Choose and use the appropriate normalisation functions available in sklearn.preprocessing and scale the data appropriately.
2. Use the Support Vector Machine algorithm to build the model.
    a. Describe SVM. Again, this is not in your lecture content, you need to do some self-learning.
    b. In SVM, there is something called the kernel. Explain what you understand from it.
    c. Write the code to build a predictive SVM model using your training dataset. (Note: You are allowed to engineer or remove features as you deem appropriate)
3. Repeat **Task A3.2.c** by using another classification algorithm such as Decision Tree or Random Forest algorithms instead of SVM.

**A4. Classification (prediction) (3 marks)**

1. Using the testing dataset you created in **Task A2.3** above, conduct the prediction for the 'GradeClass' (label) using the two models built by SVM and your other classification algorithm in **Task A3.3**.
2. Display the confusion matrices for both models (it should look like a 5x5 matrix). Unlike the lectures, where it is just a 2x2, you are now introduced to a multi-class classification problem setting.
3. Compare the performance of SVM and your other classifier and provide your justification on which one performed better.

### A5. Independent evaluation (Competition  ) (2.5 marks)

1. Read the **Student_List_A2_Submission.csv** file and use the best model you built earlier to predict the 'GradeClass for the students in this file.
2. Unlike the previous section in which you have a testing dataset where you know the 'GradeClass' and will be able to test for the accuracy, in this part, you don't have a 'GradeClass' and you have to predict it and submit the predictions along with other required submission files.
    - Output of your predictions should be submitted in a CSV file format. It should contain 2 columns: 'StudentID and 'GradeClass'. It should have a total of 162 lines (1 header, and 161 entries).
    - Hint: you may need to apply some of the data wrangling steps in A1 to this new data file (i.e., Student_List_A2_Submission.csv), to prepare it for prediction.

## Task B: Selection of Dataset, Clustering and Video Preparation

### B1. Selection of a Dataset with missing data and Clustering (4 marks)

We have demonstrated a k-means clustering algorithm in week 7. Your task in this part is to find an interesting dataset and apply k-means clustering on it using Python. For instance, Kaggle is a private company which runs data science competitions and provides a list of their publicly available datasets: https://www.kaggle.com/datasets

1. Select a suitable dataset **that contains some missing data and at least two numerical features**. Please **note** you cannot use the same data set used in the applied sessions/lectures in this unit. Please include a link to your dataset in your report. You may wish to:
    - provide the direct link to the public dataset from the internet, or
    - place the data file in your Monash student - google drive and provide its link in the submission.
2. Perform wrangling on the dataset to handle/treat the missing data and explain your procedure
3. Perform k-means clustering, choosing two numerical features in your dataset and create k clusters using Python (k>=2)
4. Visualise the data as well as the results of the k-means clustering, and describe your findings about the identified clusters.

**B2. Video Preparation (2 marks)**

Presentation is one of the important steps in a data science process. In this task you will need to prepare an **up to 3 minutes** video of yourself (you can share your code on screen) and describe your approach on the above task (**Task B1**).

- Please make sure to keep your camera on (show yourself) during recording. You may want to share your screen with your code while you talk.)

Good Luck! ☺