**Project: ML Modelling**

# Introduction

This project focuses on *data modelling*, a core step in the data science process. The task is to use the different analytical methods to solve the tasks.

# Task 1: Data Preparation and Analysis

In this assignment, you will use the csv dataset. The dataset is related to attributes and quality of wine.

The inputs include physicochemical tests (e.g. PH values) and the output is wine quality, which scales between 0 (very bad) and 10 (very excellent). There are 12 variables (i.e., attributes or columns) and 4781 instances (i.e., entries or rows). For description of attributes, check the file Readme-A2data.txt.

**Task 1.1.** Load the CSV data from the file (using appropriate Python/Pandas functions). Take a **random** sample (i.e., subset) of **600** instances from the data set, and ensure that these 600 instances don’t have any missing values. Write the random sample into a CSV file. Use this file for all subsequent tasks in this Assignment, unless otherwise specified.

**Task 1.2.** Explore the relationship between two variables: **alcohol** and **density**.

* Show the relationship in an appropriate graph (i.e., chart). Describe any interesting relationships (or lack of relationships) that you observe from the visualisation.
* Build a linear model (i.e., Simple Linear Regression) for the two variables, with alcohol being dependent variable and density as independent variable. Present the linear model in the Report and interpret the coefficients of the model.

**Task 1.3.** Explore the relationship between two variables: **quality** and **alcohol**.

* Create the side-by-side boxplot for alcohol grouped by quality level.
* Summarise your findings based on the boxplot.

# Task 2: Classification

Use the random sample data file and complete the following clustering (sub)tasks. In Task 2, you need to classify the wine quality based on the other variables.

**Task 2.1.** Select a model, either k-NN (k-Nearest Neighbours) or Decision Tree. Train and evaluate the model appropriately. Use at least 3 metrics for evaluation.

**Task 2.2.** Study the impact of at least one key parameter of the model. Describe your findings. Choose the best value(s) for the parameter(s) and justify your choice.

**Task 2.3.** With the above optimal parameter(s), train and test the model on different training/test data splits: 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20. What is the best train/test split? Why?

# Task 3: Clustering

Use the random sample data file and complete the following clustering (sub)tasks. In Task 3, you need to conduct clustering upon the random sample dataset. Don’t use the output variable **quality** when building the model or tuning the parameters (i.e., Task 3.1 or Task 3.2).

**Task 3.1.** Select a model, either k-means or DBSCAN. Build and evaluate the model. Tune the key parameter(s) of the model and justify your choice of the value(s).

**Task 3.2.** Determine the optimal number of clusters, and justify.

**Task 3.3.** Analyse the meaning (i.e. predicted quality level) of each cluster by checking the clustering results against the true quality levels (i.e., the variable **quality**). Construct and explain the confusion matrix of the results.

# Task 4: Report / Presentation

Write your report and save it in a file pdf or docx file. The quality of the report will be considered, e.g. clarity, grammar mistakes, the flow of the presentation.

Remember to clearly cite any sources (including books, research papers, etc.) that you referred to while designing aspects of your programs.