STU33011 MLA Assessment

This assignment is worth 20% of your mark for STU33011. It will consist of a report with a maximum length of ten pages. Assessments should be submitted via Turnitin by **4pm Wednesday 17th November 2021**. If you have any questions, contact me on arwhite@tcd.ie, or post a message on the class discussion board page on blackboard.

I would like you to analyse the Healthy Flow data set. This data set was collected as part of a flow cytometry analysis, and is available to download on blackboard and on the class website: https://www.scss.tcd.ie/~arwhite/Teaching/STU33011/hfd.csv.

Peripheral blood mononuclear cells (PBMC) were collected from a healthy individual. Cells were stained using labeled antibodies against CD3, CD4, CD8, and CD19 protein markers. The standardised expression levels of cells for these markers were recorded and form the first four variables in the data set.

The sample contains populations of lymphocyte cells. Each population of cells has unique characteristics with respect to the protein markers. A common goal of flow cytometry analysis is to perform “gating”, an exercise which aims to identify distinct populations of such cells within a data set. The Gate variable records the results of one such gating analysis. This variable takes values {1, 2, 3, 4}.

Your analysis should have two key objectives:

- Using unsupervised learning methods, can you identify subsets of the data which are similar to the identified gated populations?
- Using supervised learning methods, can you accurately predict which data are assigned to the identified gates?

Your report of this analysis should consist of no more than ten A4 pages (please note that this is a limit, not a target!) Please do not include any code that was used in the analysis in your report; graphs and summary output are fine. Your report should be well written and understandable to somebody unfamiliar with multivariate analysis. For example, it should be accessible to a researcher in the School of Biochemistry and Immunology.

The report will be marked subject to the following criteria. Creativity will be rewarded:

- Data description and visualisation [20%]
- Appropriate use of unsupervised learning methods [30%]
- Appropriate use of supervised learning methods [30%]
- Clarity of writing and exposition, overall report structure [20%]