**Bios/CS 534 Final Project**

In an attempt to develop blood protein biomarkers for a certain type of cancer, the level of 120 proteins were measured in the blood plasma of some patients and controls. Please download the data “project\_data.txt” from the course website:

<http://web1.sph.emory.edu/users/tyu8/534/>

The last column of the table is the clinical outcome – cancer/normal.

Please write an analysis report on the data, containing the following items:

(1) Use a dimension reduction technique to show visualization of data points in low dimensional space. Does there appear to be subgroups of proteins, and subgroups of samples? Are those subgroups associated with the clinical outcome?

(2) Use at least two clustering techniques to cluster the proteins. By comparing cluster membership of the proteins, do the two methods yield similar results?

(3) Split the data in a 3:1:1 ratio into training, validation, and testing sets. Try to construct a predictor of cancer/normal status using the proteins.

* Use at least 3 classifiers, and tune each classifier with cross-validation using the training data. At least one of the classifier should report variable importance.
* Select the best classifier based on their performance on the validation data.
* Assess the performance of the best classifier using the testing data.
* Which proteins are the top predictors?

Prepare the report as follows:

(a) Summarize the main findings, such as what methods were used and their main results, only with essential plots and tables. Structure the report with four sections: “Abstract”, “Methods”, “Results and Discussion”, “Conclusion”.

(b) Other information, such as the tuning of the methods and selection of parameters, should go into “Supplementary File 1”.

(c) All codes go into “Supplementary File 2”.

Concatenate the three into a single PDF and submit via Emory Canvas.