Assignment 02 [225 pts] **BUS Q775** (Winter 2023)
Published Date: 13/02/2023 **Due Date:** 14/04/2023

[Include all relevant codes for all questions. In addition, submit a report summarizing your models' assumptions, answers and discussion for each question. ***You are allowed to work on this assignment in pairs.*** *If you choose to do so, you are required to submit a report detailing the work that was done by each party.]*

1. **[30 pts]** Use the data `cdc.csv` uploaded in A2L to develop a model for predicting disease occurrence by location. Clearly describe your choice for the data, model, and results. Report the confusion matrix for your predictions and the area under the ROC curve.

2. **[20 pts]** Consider the Ecoli data from the machine learning repository (URL: https://archive.ics.uci.edu/ml/datasets/ecoli) and use K-means to classify it. Determine the appropriate number of clusters and plot your result to show the clusters as well as their centroids.

3. **[50 pts]** Take all the tweets by `@McMasterU` (up to the time you work on this question) as your data set and develop a classifier to classify such tweets into 25 topics: students-general, faculty-general, faculty-*i*, student-*i*, where *i* is in {Business, Engineering, Health Science, Humanities, Science, Social Sciences}, staff, community, teaching, research-*j* where *j* in {*business, fundamental, health, indigenous, materials, sustainability, social*}, other. Use both Naive Bayes and SVM methods. Describe all your steps clearly, including data gathering, cleaning and any preprocessing you have done. Report the confusion matrix for each method. Comment on your findings and report any policy insights you may draw from your classification.

4. **[50 pts]** Design a predictor for the solution of this class of knapsack optimization problems: $\{\max \ c^T\mathbf{x} \ | \ \sum_{i=1}^{n} a_i x_i \leq b \, , \ x \in \{0,1\}^n\}$. Describe all your steps and findings. Compare your predictor with a commercial solver.

5. **[100 pts]** In this question, you will solve a Kaggle competition challenge: Predict Student Performance from Game Play (URL: https://www.kaggle.com/competitions/predict-student-performance-from-game-play/overview). This competition aims to predict student performance during game-based learning in real time. You'll develop a model trained on one of the largest open datasets of game logs.

   Instruction of Accessing Data: In order to access training and testing datasets, you must need to register on Kaggle's platform and join the competition as a team. That is, if you work with another classmate, then you should join a team. After joining the competition, you can go to the **'Data'** tab to download data. You need the following three datasets for this assignment.

   - **train.csv** - the training set
   - **test.csv** - the test set
   - **train_labels.csv** - correct value for all 18 questions for each session in the training set

   Marking Rubrics:
   i. **[20 pts]** Accessing datasets and labelling training data (`train.csv`) using training_labels.csv data.
   ii. **[30 pts]** Data management, cleaning and ready for training and validation.
   iii. **[30 pts]** Train an appropriate predictive model for student performance.
   iv. **[20 pts]** Model evaluations with confusion matrix and F1 scores.