MEMORIAL UNIVERSITY OF NEWFOUNDLAND

DEPARTMENT OF MATHEMATICS AND STATISTICS

A1 **DSCI 6602** DUE: JAN. 26TH, 2023

Upload your solution to Brightspace at the beginning of class on Thursday Jan. 26th, 2023.

1. **Linear regression on the start-up dataset.** Download the start-up dataset from Brightspace and analyse it in `Python`. The goal is to do multi-linear regression to predict the profit of a start-up based on the features in the dataset.

   Your investigation should include answering the following question:

   (a) How many samples does this dataset have?

   (b) Are there any missing values in the dataset?

   (c) Is linear regression a suitable method for this dataset at all?

   Note that there are some categorical features in the dataset. These would need to be converted to numerical data first!

   For the linear regression, use 80% of the data for training and 20% for testing. Visualize the predictions and assess whether the model works well on the test data, e.g., by computing the R^2-score.

2. **Logistic regression on the Titanic dataset.** Download the Titanic dataset from Brightspace and analyse it in `Python`. The goal is to predict the probability of survival of the passenger as a function of gender, class, etc.

   To complete this task, answer the following questions:

   (a) Are there any missing values in the dataset? If so, what would be the most appropriate way of dealing with them?

   (b) Have there been more male or more female survivors? What about the influence on which class the passengers travelled in?

   (c) Which features of the dataset do you think need to be included for predicting the survival of a passenger (to answer this question quantitatively, you could train various models with a subset of the given features)?

   What is the final accuracy of your model, again using a 80%/20% split for training/test data. Are there any other verification measures you think would be meaningful here (e.g. precision, recall, etc.)?