MIS40970 Assignment 3
Classification

Submission Details & Deadline
Please submit a report (pdf) through Blackboard by 20h30 Wednesday 19 April. This assignment is worth 20% of the total credit for MIS40970. In your submission, please make sure that in each case you include any code or images of screenshots which were used to address each item below. Also, include the output (if any) generated, i.e., Show Your Work!

Datasets
The *colleges.csv* dataset is available on Blackboard. The *churn* and *GermanCredit* datasets are included in R packages "C50" and "caret".

Assignment
Apply what you have learned about Data Mining to date. Include plain English interpretation of your analyses.

Q1 Compare and contrast classification and clustering. [10 marks]

Q2 Describe what this piece of R code is doing and why it is an important starting point for running classification algorithm. [10 marks]

```
> set.seed(1234)
> dataPartition <- sample(2,nrow(data),replace=TRUE,prob=c(0.7,0.3))
> trainData <- data[dataPartition ==1,]
> testData <- [dataPartition ==2,]
```

Q3 What is the role of the M parameter in the Weka implementation of C4.5 algorithm? Which part of the DTL induction process does this parameter affect? [5 marks]

Q4 Install R package "C50". Import customer churn dataset (churn) using *data()* function. Examine the *churnTrain* dataset. Using R run a decision-tree classification algorithm of your choice constructing a full unpruned tree and a pruned tree. Compare classification results of the pruned and unpruned trees generated. [marks 10]

Q5 Compare generalisation performance of the pruned and unpruned tree from Q4. Output relevant summaries and confusion matrices. Describe the results. [10 marks]

Q6 Install R package "caret". Import German credit rating dataset (*GermanCredit*). Examine the data. Use the data to build a classification model to predict "Good" or "Bad" customer credit rating. Pay attention to the model's generalisation and its' ability to correctly predict both classes. Interpret the results. [15 marks]

Q7 Load file *college.csv* provided on Blackboard. Explore the data. Prepare the data for analysis. Use three different classification algorithms to classify colleges into two classes based on the label ("Not Elite", "Elite") (at least one algorithm of the type decision tree).

1. Describe what you have learned about the dataset and classification results.
2. What classification algorithm(s) did you use, with what parameter settings and how these settings affected the algorithm(s) performance?
3. Exclude the class labels from the data and explore this dataset with clustering. Compare clustering results with results of classification.

[40 marks]