**ST4003/CS7DS1 Data Analytics Assignment**

This assignment has two parts corresponding to 2 different datasets. The first dataset is the fish dataset which Michael Keating spoke about during the week and the second dataset relates to employee attrition. You are required to write report for each of the model building exercises explaining the output. It is worth 40% of the total marks for this module.

Data Set 1

This has two files main files (**economic**) which includes economic variables per fishing boat while the second file has data on the types of fish(**fish2**) with corresponding monetary value caught for each boat across the years. The aim of the assignment is to build a model to predict net profit. (You are welcome to pick another target variable if you like). There is an additional file called **species** which documents the species.

You should include

* An examination of the data.
* A single tree model
* Two ensemble techniques of your choice
* Comparison of the above models

We are going to consider the cases across years as independent cases. There is a year variable included in the data. .

For the examination of the data marks will be awarded for good presentation. Pictures of R output are not acceptable throughout the report. Obviously you should look at all variables. Your report should include a detailed examination of 3 variables of your choice. Please include at least 1 categorical and 1 continuous. .

For each of the models built you should investigate changing the parameters inherent in the technique. All models should be evaluated and the additional output from each model should be discussed. Marks will be awarded for good presentation throughout this section.

Marks will be awarded as follows:

* An examination of the data 10%
* **A single tree model 15%**
* **Ensemble Technique 1 25%**
* **Ensemble Technique 2 25%**
* **Comparison of the above models 15%**
* **Creativity 10%**

Data set 2

This data relates to Employee attrition. It is an Excel file(**Employee attrition**) with the first sheet showing the data while the second sheet describes some of the variables.

The aim of the assignment is to predict who is going to leave the organisation (I.e. Attrition = Yes). You should include

* An examination of the data.
* A single tree model
* Two ensemble techniques(different techniques than for the previous dataset)
* Comparison of the above models

For the examination of the data marks will be awarded for good presentation. Pictures of R output are not acceptable throughout the report. Obviously you should look at all variables. Your report should include a detailed examination of 3 variables of your choice. Please include at least 1 categorical and 1 continuous. .

For each of the models built you should investigate changing the parameters inherent in the technique. All models should be evaluated and the additional output from each model should be discussed. Marks will be awarded for good presentation throughout this section.

Marks will be awarded as follows:

* An examination of the data 10%
* **A single tree model 15%**
* **Ensemble Technique 1 25%**
* **Ensemble Technique 2 25%**
* **Comparison of the above models 15%**
* **Creativity 10%**

The assignment is to be submitted via Turnitin. The Class ID is 16791777 and the enrolment key is Trinity. You can only submit an assignment once. Please hand in a hard copy of your report to jean Maypother by Monday 18th December at 12.00. To avoid confusion please do not wait till the 12.00 deadline.