Reading Topic 3: One or more of the following (or equivalent).

- Venables, W.N. and Ripley, B.D. (2002). Modern Applied Statistics with S, Springer. Chapters: 9
- Rpart manuals:
  `https://cran.r-project.org/web/packages/rpart/rpart.pdf`
  `https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf`
- ROCR manual & other materials: `https://rocr.bioinf.mpi-sb.mpg.de/`
- Fawcett, T. (2006). An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861874. [see webcourses module].

Assignment 4: For 10% module credit (MSc group);
25% Module credit (PhD group)
Submission deadline: 6.30pm Monday 24th April - hard-copy only!

The student_retention_full.csv dataset is available on Webcourses. **Please ensure you are using the correct dataset!**

This analysis is concerned with data relating to student retention in the Engineering faculty of DIT. It is the same data considered in Assignment 3, but with more potential predictors available. Details on the variables are given in Table 1 below.

The purpose of the analysis will be to use Rpart models to build a predictive model for student retention. The data includes risk risk factors regarding prior academic performance (e.g. leaving certificate results, leaving certificate maths grade), personal characteristics (gender, home address, CAO choices made etc).

1. Write an R programme to read in the student_retention_full.csv dataset and analyse these data. The response is whether or not that a student proceeded to second year of their programme or not. (i.e. overall6=1 or 0) and you should initially fit an rpart tree model including all the other variables as predictors.

   [20 marks]

2. You should suitably prune this tree and calculate performance criteria (e.g. ROC analysis, AUC etc.).

   [20 marks]

3. Present a data analysis report (approx. 5 pages): briefly outlining the tree building algorithm and validation analysis; report on the fitted models (include appropriate plots) and your findings.

   [60 marks]

4. Submit both this report and the **R** programme code you used for the analysis.

Variable name	Description	Variable Format
ID	Unique observation identifier	Number (1 to 607)
Response variables		
Overall6	Overall First Year Engineering Result	1 = Student progressed to second year 2 = Student did not progress to second year or withdrew from course
Predictor variables		
Gender	Gender of individual	0 = Female 1 = Male
Address2	Home address	1 = Dublin 2 = Greater Dublin Area 3 = Rest of Ireland
Caochc2	Central Application Office choice	1 = 1st preference 2 = 2nd preference 3 = 3rd preference 4 = 4th preference
Lcpoints (Lcgradefinal)	Total Leaving Certificate points	Number 0 to 600
Lcmaths (lcmthpts)	Leaving Certificate points for Mathematics.	Number 0 100
Mathlev1	The level of mathematics taken at Leaving Certificate.	1 = Ordinary 2 = Higher
Lcphysic	Leaving Certificate points for Physics.	Number 0 80
Havephy	Indication of having studied not studied Leaving Certificate Physics	0 = Has not studied LC physics 1 = Has studied LC physics
Lcchem	Leaving Certificate points for Chemistry.	Number 0 80
Havechem	Indication of having studied not studied Leaving Certificate Chemistry	0 = Has not studied LC chemistry 1 = Has studied LC chemistry
Lcenr	Leaving Certificate points for Engineering.	Number 0 80

Variable name	Description	Variable Format
Haveenr	Indication of having studied not studied Leaving Certificate Engineering	0 = Has not studied LC engineering 1 = Has studied LC engineering
Lctdr	Leaving Certificate points for Technical Drawing.	Number 0 80
Havetdr	Indication of having studied not studied Leaving Certificate Technical Drawing	0 = Has not studied LC Technical Drawing 1 = Has studied LC Technical Drawing
Lcpch	Leaving Certificate points for the subject Physics-Chemistry	Number 0 80
Havepch	Indication of having studied not studied Leaving Certificate Physics-Chemistry	0 = Has not studied LC Physics-Chemistry 1 = Has studied LC Physics-Chemistry
Lcappm	Leaving Certificate points for Applied Mathematics.	Number 0 80
Haveappm	Indication of having studied not studied Leaving Certificate Applied Mathematics	0 = Has not studied LC Applied Mathematics 1 = Has studied LC Applied Mathematics
CAOCHC2	Course choice under the CAO application. 1st,2nd, etc	1 = First 2 = Second 3 = Third 4 = Fourth
CORE1	Core subject 1 studied as part of the 1st year.	Subject e.g. Mathematics
CORE2	Core subject 2 studied as part of the 1st year.	Subject e.g. Mechanics
CORE3	Core subject 3 studied as part of the 1st year.	Subject e.g. Physics
AWARD	Award level of course	1 = Higher 2 = Ordinary

Table 1: dataset variables.