# CAB330 – Data and Web Analytics

**Task 1) Data Selection and Distribution. (4 marks)**

1. What proportion of households who have high risk?
2. Did you have to fix any data quality problems? Detail them?
3. The dataset may include irrelevant and redundant variables. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
4. What distribution scheme did you use? What “data partitioning allocation” did you set? Explain your selection. (Hint: Take the lead from Week 2 lecture on data distribution)

**Task 2) Predictive Modelling Using Decision Trees (4 marks)**

1. Build a decision tree using the default setting.
   1. What is classification accuracy on training and test datasets?
   2. Which variable is used for the first split? What are the variables that are used for the second split?
   3. What are the 5 important variables in building the tree?
   4. Report if you see any evidence of model overfitting.
2. Build another decision tree tuned with GridSearchCV
   1. What is classification accuracy on training and test datasets?
   2. What are the parameters used? Explain your decision
   3. What are the optimal parameters for this decision tree?
   4. Which variable is used for the first split? What are the variables that are used for the second split?
   5. What are the 5 important variables in building the tree?
   6. Report if you see any evidence of model overfitting.
3. What is the significant difference do you see between these two decision tree models – default (Task 2.1) and using GridSearchCV (Task 2.2)? How do they compare performance-wise? Explain why those changes may have happened.
4. From the better model, can you identify which householders to target for providing loan? Can you provide some descriptive summary of those householders?