BSc in Software Design - C.A (40%)

Choose data sets from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) or elsewhere. Include a link in your document to a documentation page for the datasets.

Submit a python folder including

- Two notebooks
- Plots in a plots folder
- Data sets in a "data" folder
- A report which documents your analyses as a PDF file in a doc folder.

Also submit the same report as a Turnitin report. Each analysis should follow the CrispDM data mining process and include the following steps.

	Regression	DT
• Business Understanding/Objective	3	3
• Data Understanding/Data Exploration	5	5
• Modeling	5	5
• Evaluation	4	4
• Conclusion	3	3

1. Regression (20 marks)

Choose a suitable Regression data set and carry out a linear regression analysis.
(Sections 1.1 to 1.5 in the report)

2. Decision Trees (20 marks)

Choose a suitable classification data set and carry out a decision tree analysis.
(Sections 2.1 to 2.5 in the report).

Report length is a max of 1500 words and a max of 10 pages.
Longer reports will loose marks.

More on the sections

- Objective
  - Outline the objective of the data mining prediction.
  - Comment on and identify a use case for the predictions
- Data Exploration
  - Comment on all plots and analyses. Draw conclusions about the data set and whether it is likely that a good prediction model can be created.
- Modeling
  - Comment on the models. Explain your approach if you are doing cross-validation.
- Evaluation
  - Evaluate the model on the test data.
- Conclusion
  - Give a brief overview of how successful the analysis was and how useful the model will be

Notes:

- Use the section numbers as described above.
- The prediction task should be sensible. Also relate it to a business objective.
- **The report should contain snippets of code showing how models were built, etc.**
- **The report should contain results (output from python notebooks)**
-
- Use a separate notebook for each analysis. Two notebooks only.
- Don't use data sets we have used in class.
- Use relative path names for data sets and plots so code executes on other machines.
- Put a link to the dataset documentation at the start of the report.
- Include the csv files in the data folder.
- No need for a contents page in the report.
- The report should not read completely like an essay. Use lists for variable descriptions etc.
- Don't explain theory, for example what a decision tree is.
- Read in the file from the data folder. Don't use a file chooser.
- File names must be case sensitive, so read_csv() statement works on Linux
- Set a random_state in train_test_split so I can replicate results.
- **Overall structure and readability of report is important.**