Term Assignment (40%) - 2018

Choose data sets from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) or elsewhere. You will need at least two data sets, one for regression and one for classification. You will carry out three analyses.

For each data set you should

1. Identify the problem to be solved
2. Explore the data (producing tables and graphs)
3. If necessary set aside some data for testing purposes (define training set and testing set)
4. Create a prediction model
5. Apply the model to the test data and evaluate.
6. Evaluate the results.

You should apply three data mining techniques.

- Regression
- Decision Trees
- kNN

Write a report. The report should consist of three parts, one for each technique. For each technique you should have the following sections.

1. Problem Identification and Overview of the data sets
2. Data exploration
3. Definition of training and testing data
4. Model Generation
5. Predictions & Evaluations for the test data
6. Conclusion

Section numbers run from 1, 1.1, 1.2 ..... 3.6
If you choose the same data set for two techniques, then there is no need to repeat subsections 3.2. However section 3.6 should then include a comparison between the results for both techniques.

Note that
1. Overview of the data sets should not be a cut and paste from UCI repository. You can reference these descriptions.
2. Graphs are only considered useful if you comment on them and come up with some conclusions/insights.
3. Give snippets of code for definition of training and test data.
4. Model Generation section should include code snippets and comments on the model.
5. For regression, evaluation for test data should include the RMSE for the training data.
6. Conclusion can address model itself and the performance on the test data.

Submit a single zipped R project (folder). The name of the zip file, folder and R project should be your student ID.

Include

- Data sets in a “data” folder
- R scripts
- Report as a PDF file in doc folder.