ML 4375 – Intro to Machine Learning – Summer 2018 – Mazidi

Project 1

* Select 2 data sets for the project. One data set should be suitable for regression and the other for classification.
* For *each* data set:
  + Decide which column is the target: what you want to learn.
  + Use at least 5 R functions for examining each data set (head, summary, names, cor, str, etc.)
* Produce at least 2 informative graphs for each data set.
* Run *at least* 2 ML algorithms on each data set, using at least 3 algorithms in all.
  + Run summary() on each model.
  + Select an appropriate metric to evaluate how well each algorithm performed on the test set. Provide some commentary and analysis of which algorithm did best and why you think that is the case.
  + Algorithms we have learned so far:
    - regression: linear regression
    - classification: logistic regression
    - regression or classification: knn
* Upload the Rmd file to eLearning.
* Part of your grade (10 points) is how well you graded your peers.
* Peer Evaluation
  + see Peer Review file
  + We will randomly assign 3 peers to evaluate your work. These scores will be averaged by the TA/instructor. The TA/instructor reserves the right to raise or lower the grade in the event of poor grading by your peers.

Top and/or interesting projects will be chosen by the instructors to present to class.

Sources for datasets:

* <https://www.kaggle.com/datasets> (free but you have to join)
* <http://archive.ics.uci.edu/ml/>
* <https://blog.bigml.com/list-of-public-data-sources-fit-for-machine-learning/>