# Machine Learning Coursework Notes

Propose doing Decision Tree compared to Random Forest

## Datasets

### Letter Image Recognition Data

* <http://archive.ics.uci.edu/ml/datasets/Letter+Recognition>
* 16 features describing letters in various fonts and distortions
* 1 target (Letter), 26 classes
* 20 000 examples
* Poor papers
* Clean dataset, balanced
* Default tree gives approx. 60% predictability, papers gives 80%

### MAGIC gamma telescope data 2004

* <http://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope>
* The data set was generated by a Monte Carlo program to simulate extensive air showers of atmospheric gamma radiation. 10 features, 1 target (G / H)
* 19 200 examples
* Good papers
* Clean dataset, unbalanced (12332 Background Gamma/ 6688 Hadron noise)
* Good results with default dataset

To Aaron:

This is an optional step. However, if there is a significant class imbalance in your dataset, it's reasonable to apply oversampling/undersampling during cross validation (only on training set, not validation samples). When you compare algorithms, the two models have to be retrained on the entire training set with e.g. SMOTE using the optimal hyper-parameters discovered during model selection. Yes, you can present the best performing one. To enrich your analysis, optionally you could investigate the impact of varying degrees of SMOTE on our final models.