IN3062 – Report

• Define the problem domain and dataset(s) (you are free to choose the problem domain and the dataset that you want to investigate).

• Define questions and analysis tasks (a brief overview of the domain, the questions that

are being asked, a list of your objectives and the expected output(s) of your analysis).

• Perform an initial investigation of the dataset and the characteristics of the data. Develop

a plan as to how you might transform the data to make it useable.

• Develop a plan as to which artificial intelligence techniques you might use and what sorts

of potential observations these can lead to, and how you will evaluate these.

• Use models taught in the module. You must use models taught in the module, these are:

perceptron, decision trees, linear regression, support vector machines, random forest, k-

nearest neighbour, naïve Bayes, neural networks as well as unsupervised techniques k-

means and GMM, and principal component analysis. Most supervised models have both

classification and regression variants. You are encouraged to work with neural networks.

An additional technique from outside the taught module content might be applied for

comparison purposes, if this is done it should be clearly indicated and well justified.

• Split your dataset (train/validate/test, some datasets come pre-split). If you have a holdout

test set then you most likely don’t want to use this until the near the end of your work.

• Perform the analysis. Get the data ready for analysis, carry out your analysis/modelling

as needed, validate your results and communicate observations, iterating through this

process. Analytical operations can include data processing to an extent that is needed

(not all datasets are messy) to prepare a useful and robust dataset to work within, and data

derivation (such as feature engineering).

• You might establish a baseline result first, computing metrics on training and validation

sets, analyse errors, work on succeeding iterations, and alternative models. (If initial

baseline results are amazing and there are no errors is the problem too easy?)

• Generally, be close to your data (visualise the dataset, collect summary statistics, look at

errors, analyse how different parameters affect performance, try out different model

variants)

What is your dataset, problem domain?

• Is your problem classification or regression?

• Did you have any missing, corrupt or misleading data? If so, how did you cope it?

• Have you omitted some data? If so, why?

• Did you apply techniques to understand your dataset?

• How did you encode the input variables?

• What models did you use?

• What are the criteria for selecting model performance evaluation tools?

• What were your outputs?

• Did you have any problems or difficulties working with the dataset?

**Report, introduction: description and motivation of the problem, description of the**

**dataset including data types (e.g. discrete, continuous) (15%)**

**Report, methodology: summary of the models used, with their pros and cons, a**

**hypothesis statement, description of choice of training and evaluation methodology**

**(20%)**

**Report, results: description and presentation of the output. The code acts as an**

**appendix to this section, and code quality (e.g. commenting) contributes. (30%)**

**Report, evaluation: analysis and critical evaluation of results. (10%)**

**Reflection and discussion (20%)**

**Conclusion**

**References**

references (using Harvard format) and future work. (5%)