We will be using the WBCD (Wisconsin breast cancer database) data set for this assignment, and some other assignment that will follow.

This data set is a very commonly used dataset for comparing the performance of algorithms (see attached paper).

Download the data set from the below url, or do a Google search for Breast Cancer Wisconsin dataset <http://mlr.cs.umass.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)>. Pre-processing – remove data samples with missing values, as done in the research paper. Use the same train-test division as in the research paper. Classify using logistic regression. Now get the original data set. Instead of removing the samples with missing values, replace the missing values with 1. Use a train-test division comparable to what you used above.

Logistic regression can be done even using MS Excel. However, you are required to implement this using some programming language, s.a. python.

Deliverables:

Implemented algorithm.

A small report (max 1 pg) – discuss how the algorithm was implemented, and the results you obtained.