The breast cancer is comprised of a total of 699 patient's records with cancer diagnosis information. The dataset is derived from cell images from which cell-specific features are extracted. Each data contains 10 features, namely:

radius (mean of distances from center to points on the perimeter)

texture (standard deviation of gray-scale values)

perimeter

area

smoothness (local variation in radius lengths)

compactness (perimeter^2 / area - 1.0)

concavity (severity of concave portions of the contour)

concave points (number of concave portions of the contour)

symmetry

fractal dimension ("coastline approximation" – 1)

Training the classifier parameters (or classifier model) involves passing the training

data with ground-truth label information to the supervised classifier learning module.

This is achieved by executing the following command:

$hadoop jar /usr/lib/hadoop/Test.jar com.winvector.logistic.demo.MapReduceLogisticTrain

input/WCTrng.tsv "DIAGNOSIS ~ TEXTURE + PERIMETER + AREA + SMOOTHNESS + COMPACTNESS + CONCAVITY + CONCAVEPOINTS + SYMMETRY + FRACTALDIMENSION" BC\_model.ser

Validation of Classification with Test Data

hadoop jar /usr/lib/hadoop/Test.jar com.winvector.logistic.demo.MapReduceScore

BC\_model.ser input/WCTest.tsv BCScore

Where, input/WCTest.tsv – Test data set

BCScore – output directory (in HDFS)