The problem is to determine whether a patient referred to the clinic is hypothyroid. Therefore three classes (label) are built: normal (not hypothyroid, hyper function and

subnormal functioning). Because 92 percent of the patients are not hyperthyroid a good

classifier must be significant better than 92%.

Data Set up Summary:

Number of attributes: 21 (15 attributes are binary, 6 attributes are continuous)

Number of classes (label): 3

Number of learning examples: 3772

Number of testing examples: 3428

Generate descriptor for the dataset using the command:

bin/hadoop jar $MAHOUT\_HOME/mahout-core-0.6-job.jar

org.apache.mahout.classifier.df.tools.Describe -p Thyroid/ann\_train.arff -f

Thyroid/ann\_desc.info -d 21 N L

Generate the model using command:

bin/hadoop jar $MAHOUT\_HOME/mahout-examples-0.6-job.jar

org.apache.mahout.classifier.df.mapreduce.BuildForest -

Dmapred.max.split.size=1874231 -d Thyroid/ann\_train.arff -ds Thyroid/ann\_desc.info

-sl 5 -t 5 -o Thyroid/model

Classify the test data using command:

bin/hadoop jar $MAHOUT\_HOME/mahout-examples-0.6-job.jar

org.apache.mahout.classifier.df.mapreduce.TestForest -i Thyroid/ann\_test.arff -ds

Thyroid/ann\_desc.info -m Thyroid/model -a -mr -o output-Thyroid