Name – Kanika Kapoor

NetID – kxk140230

Date of Submission – September 25, 2015

Files:

1) kxk140230_kyphosis1 – R script that contains kyphosis data split as Training Data – 80% and Test Data – 20%.

2) kxk140230_kyphosis2 - R script that contains kyphosis data split as Training Data – 90% and Test Data – 10%.

3) kxk140230_solder1 - R script that contains solder data split as Training Data – 80% and Test Data – 20%.

4) kxk140230_solder2 - R script that contains solder data split as Training Data – 90% and Test Data – 10%.

Results Table:

DATASET NAME	PARTITION	ACCURACY OF TEST DATA WHEN PRUNED MODEL OF TRAINING SET IS USED FOR PREDICTION	ACCURACY OF TEST DATA WHEN NORMAL TRAINING SET IS USED FOR PREDICTION
kyphosis	Training – 80% Test – 20%	0.8235294	0.7647059
kyphosis	Training – 90% Test – 10%	0.78189	0.7777778
solder	Training – 80% Test – 20%	0.625	0.5833333
Solder	Training – 90% Test – 10%	0.59681	0.5555556

Observations I made from Results table:

1) Accuracy is best when data is split 80-20 and prediction is based on Pruned Training Set.
2) Accuracy drops when we use just the training set for prediction.
3) In case of kyphosis where data(n=81) is not very big and split into 90-10, accuracy is almost same when prediction is done using pruned training set and normal training set.

Reference:

- http://michael.hahsler.net/SMU/EMIS7332/R/chap4.R
- http://www.stat.pitt.edu/sungkyu/course/2221Spring15/rpart.txt