Name – Kanika Kapoor

NetID - kxk140230

Title – Machine Learning Assignment 02 Decision Trees

### **ACCURACY RESULTS**

Accuracy results have been tested for all cp values against minsplit = 2 and minsplit = 100. These 2 minsplit values have been choosen randomly whereas the default minsplit value is 20.

# 1) DATASET - 1

DATASET	MINSPLIT	VALUE OF CP	ACCURACY(%)
Dataset - 1	2	0.172308	50
		0.042308	57.1
		0.015385	61.25
		0.014615	67.75
		0.013846	69.75
		0.013462	69.75
		0.011538	71.9
		0.010000	71.9

DATASET	MINSPLIT	VALUE OF CP	ACCURACY(%)
Dataset - 1	100	0.172308	50
		0.042308	57.1
		0.015385	61.25
		0.014615	67.65
		0.013846	68.75
		0.013462	68.75
		0.011538	70.9
		0.010000	70.65

# 2) DATASET - 2

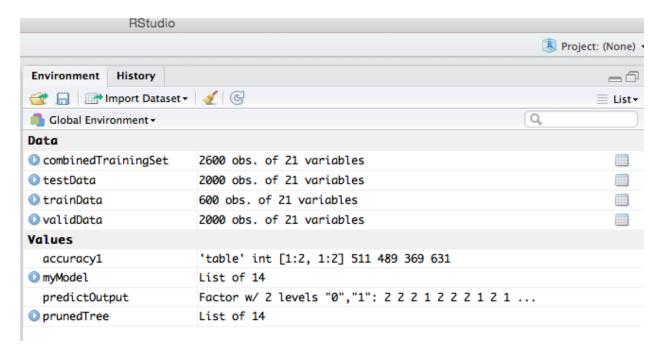
DATASET	MINSPLIT	VALUE OF CP	ACCURACY(%)
Dataset - 2	2	0.170000	50
		0.050833	61
		0.033333	62
		0.031667	62
		0.030000	70.5
		0.029167	70.5
		0.016667	70.833
		0.013333	76
		0.010000	76

DATASET	MINSPLIT	VALUE OF CP	ACCURACY(%)
Dataset - 2	100	0.170000	50
		0.050833	61
		0.033333	62
		0.031667	62
		0.030000	70.5
		0.029167	70.5
		0.016667	69
		0.013333	72
		0.010000	72

## **Workspace Screenshot**

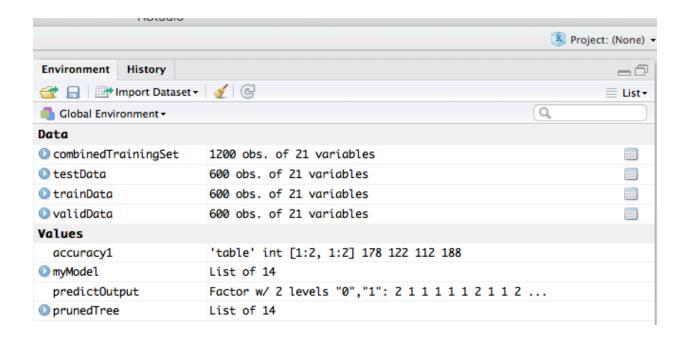
#### 1) DataSet-1

It shows that for dataset – 1, Training file contains 600 rows or observations Validation file contains 2000 rows or observations Combined Training Set contains 2600 rows



#### 2) DataSet-2

It shows that for dataset – 2, Training file contains 600 rows or observations Validation file contains 600 rows or observations Combined Training Set contains 1200 rows



#### PACKAGE AND PARAMETERS USED

Package used – rpart

It is used for recursive partitioning of decision tress for classification and regression procedures. It creates a model using parameters such as split type and minsplit value. In our case we have used two values of minsplit, 2 being the least and 100 being the max.

For pruning the dataset, it uses parameters such as cp value which is the complexity parameter. We plot a cp table using command printcp(modelname) and then find the cp value corresponding to least xerror value. We use that cp value for pruning our dataset. Once we have the pruned tree, we predict it using again the testdata. It uses parameter such as type, which can be prob, vector etc. We have used type="class" for our analysis.

All plotting is done using the plot and text functions and the results have been stored as .jpg files. The summary of data models have been saved as .txt files.

#### Reference:

https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf