Assignment 1

Sumedh Masulkar(11736)

Threshold on decrease in impurity

Impurity Function	Threshold (cp value)	xerror	Cross Validation Error	Accuracy
Information	0.005	0.72015	0.25130	74.87%
Information	0.010	0.72761	0.25390	74.61%
Information	0.015	0.67910	0.23697	76.30%
Information	0.020	0.73507	0.25651	74.34%
Gini	0.005	0.74254	0.25911	74.08%
Gini	0.010	0.73134	0.25520	74.48%
Gini	0.015	0.69403	0.24218	75.78%
Gini	0.020	0.75000	0.26172	73.82%

Root node error = 0.34896

Threshold on number of data vectors

Impurity Function	Threshold (minbucket value)	xerror	Cross Validation Error	Accuracy
Information	5	0.72015	0.25130	74.87%
Information	10	0.75373	0.26302	73.69%
Information	15	0.75000	0.26172	73.82%
Information	20	0.72169	0.25184	74.81%
Gini	5	0.75373	0.26302	73.69%
Gini	10	0.71269	0.24870	75.13%
Gini	15	0.72761	0.25390	74.61%
Gini	20	0.67910	0.23697	76.30%

Grow and prune

Impurity Function	xerror	Cross validation error	Accuracy
Gini(full grown)	0.69030	0.24088	75.91%
Gini(pruned)	0.69030	0.24088	75.91%
Information (full grown)	0.72388	0.25260	74.74%
Information (pruned)	0.68657	0.23958	76.04%

For missing values, I have used median since using surrogacy would require lot of computation and time if there are a lot of missing values, as in case of insulin and triceps. For rest of the data, surrogacy is used.