CS481 DATA SCIENCE

Lab 5 Decision Trees
February 23, 2020

Muhammad Bilal Khan 16K3778

Decision Trees Results

10 Fold Corss Validation

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
DT using gini (without pruning)	93.93	89.1	85.71	89.28	75.00
DT using gini (with pruning)	95.38	31.75	100.00	82.14	84.21
DT using entropy (without pruning)	93.84	90.05	90.47	89.28	78.94
DT using entropy (with pruning)	95.38	61.45	100.00	85.71	73.68
Standard Deviation	0.863	27.61	7.14	3.418	4.729

70/30 Hold Out Approach

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
DT using gini (without pruning)	87.24	87.31	84.12	78.57	59.32
DT using gini (with pruning)	91.32	32.03	80.95	77.38	67.79
DT using entropy (without pruning)	87.24	87.86	85.71	77.38	64.40
DT using entropy (with pruning)	90.30	60.85	85.71	73.80	69.49
Standard Deviation	2.102	26.510	2.243	2.065	4.483

K Fold Cross Validation vs. Hold Out Approach

Regarding the comparision between CV and hold out, *Cross Validation proves to be superior for smaller datasets* due to its results providing summarized results with comparision to all the dataset being divided in to training and testing set eventually, but due to comparision of K blocks with the rest of the data complexity becomes far more wide spread.
Hold out approach is considered better for bigger datasets, since it constantly divides the data in to a considerable bigger chunk of usually 70/30 for training and testing respectively.