MACHINE LEARNING
HOMEWORK – 1
Decision Trees

Vivek Anand Sampath
vxs135130

RUNNING THE PROGRAM:
$ javac decision_tree.java
$ java decision_tree 1000 30 "<traning_file_path>" "<validation_file_path>" "<test_file_path>" yes

Total parameters to be passed: 6
The program doesn't run without these 6 parameters
Parameter 1: L-value
Parameter 2: K-value
Parameter 3: path for the training set file within double quotes
Parameter 4: path for the validation set file within double quotes
Parameter 5: path for the test set file within double quotes
Parameter 6: yes | no, for printing the decision tree after pruning

IMPLEMENTATION DESIGN:
Reading the input .csv file:
1. Identify the attributes (features)
2. Identify the target attribute
3. Get the data set loaded into a 2D array

Building decision tree:
1. Decision Tree is built using the ID3 algorithm recursively by identifying the best classifying attribute having the maximum gain.
2. The best attibute among the list of attributes is selected using one of the two heuristics
    1) entropy and
    2) variance impurity

Calculating accuracy:
1. Test the decision tree constucted with the test data by tracing the attributes for each instance and the leaf node it leads to. Compare the resultant node value with the class attribute value given. Count the number of successful classifications and find the accuracy by dividing with the total number of instances.

Post-pruning:
1. Calculate the accuracy of the decision tree constructed against the validation set
2. A random non-leaf node from the decision tree is selected and replaced with leaf node assigning the most common value at that node.
3. Do this replaced till n times, where n is any random number between 1 to k(input parameter).
4. Calculate the accuracy of the tree formed as a result of this pruning, if that performs well on the validation set than the previous tree then remember the tree.
5. Repeat the pruning process l(input parameter) times and return the best tree found in terms of accuracy against the validation set

MACHINE LEARNING
HOMEWORK – 1
Decision Trees

Vivek Anand Sampath
vxs135130

Final accuracy:
1. Test the accuracy of the decision tree after pruning against the test data set.

RESULTS:

DATA SET 1:

L	K	ACCURACY (in %)			
		ENTROPY		VARIANCE IMPURITY	
		Before pruning	After pruning	Before pruning	After pruning
100	20	75.85	76.4999999	76.64999	77.10
500	30	75.85	77.2	76.64999	77.85
1000	50	75.85	76.14	76.64999	77.3
2000	50	75.85	77.10	76.64999	77.45
1000	70	75.85	76.75	76.64999	78.5
1000	80	75.85	77.60	76.64999	77.3
2000	80	75.85	76.95	76.64999	77.2
200	10	75.85	76.25	76.64999	78.05
300	25	75.85	76.4999	76.64999	76.9

MACHINE LEARNING
HOMEWORK – 1
Decision Trees

Vivek Anand Sampath
vxs135130

DATA SET 2:

L	K	ACCURACY (in %)			
		ENTROPY		VARIANCE IMPURITY	
		Before pruning	After pruning	Before pruning	After pruning
100	20	72.333	74.167	72.5	75.167
500	30	72.333	72.667	72.5	74.667
1000	50	72.333	73.83	72.5	76.167
2000	50	72.333	74.83	72.5	76.33
1000	70	72.333	73.167	72.5	74.333
1000	80	72.333	72.667	72.5	72.5
2000	80	72.333	73.167	72.5	74.83
200	10	72.333	75	72.5	72.0
300	25	72.333	74.167	72.5	74.167