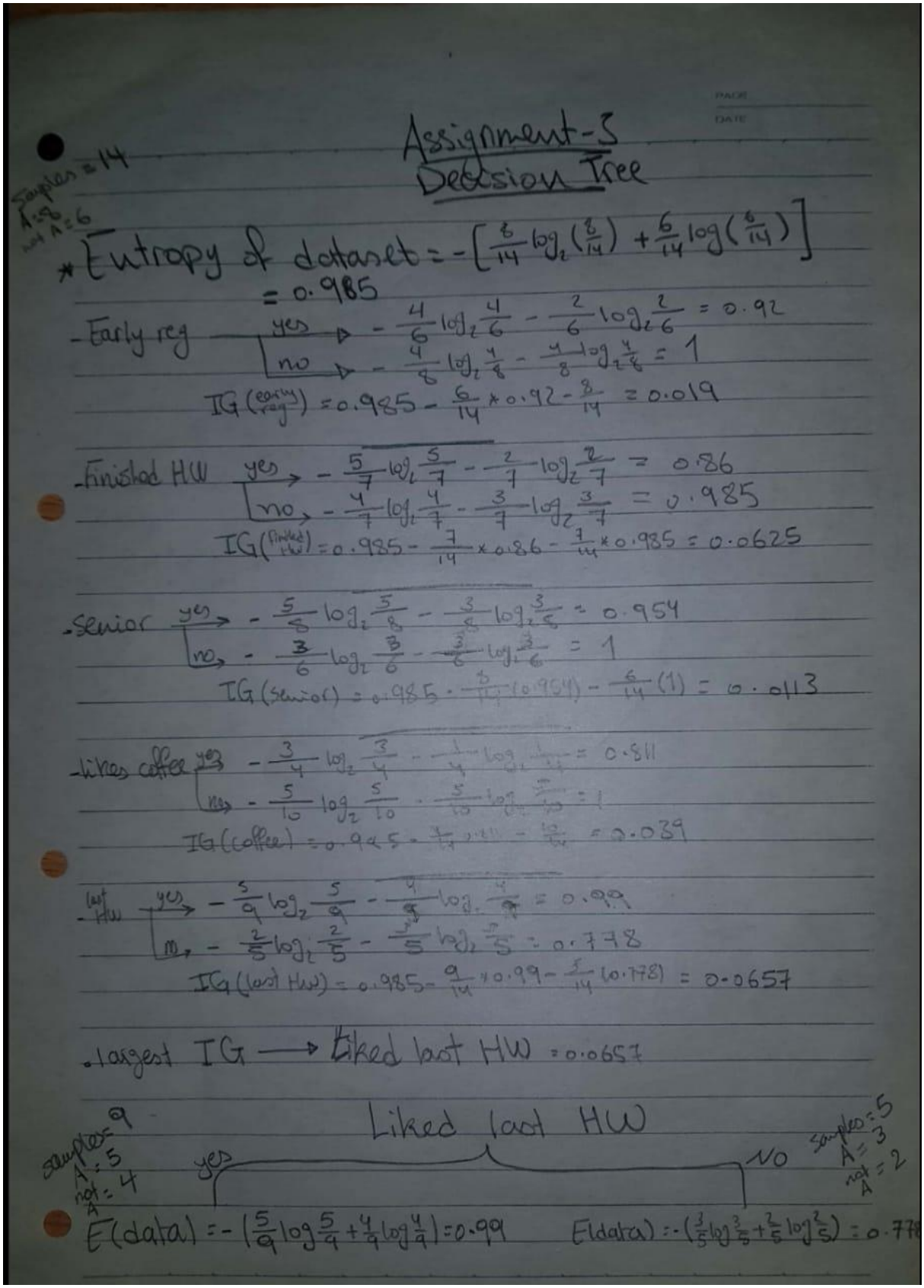
Faculty of Engineering
Cairo University

Assignment 3

Mohamed Sameh Mohamed Zaki

1170556

Manually Built Tree:



Early reg	finished HW	senior	likes coffee	A
1	1	0	0	1
1	1	1	0	1
0	1	1	0	0
0	0	1	1	1
1	0	0	0	0
0	1	0	1	1
0	0	1	0	1
0	1	1	1	0
0	0	0	0	0

E(data) = 0.99

early reg: yes → $-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.912$; no → $-\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1$

$IG = 0.99 - \frac{3}{9}(0.92) - \frac{6}{9}(1) = 0.017$

finished HW: yes → $-\frac{4}{5}\log\frac{4}{5} - \frac{1}{5}\log\frac{1}{5} = 0.72$; no → $-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$

$IG = 0.99 - \frac{5}{9}(0.72) - \frac{4}{9} = 0.15$

senior: yes → $-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.97$; no → $-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$

$IG = 0.99 - \frac{5}{9}(0.97) - \frac{4}{9} = 0.006$

likes coffee: yes → $-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$; no → $-\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1$

$IG = 0.99 - \frac{3}{9}(0.92) - \frac{6}{9} = 0.017$

⇒ finished HW has the highest IG

early reg	HW ?	senior	likes coffee	A
0	0	1	0	0
0	1	1	0	1
1	0	0	0	0
1	1	1	0	1
1	0	0	1	1

E(data) = 0.778

early reg: yes → $-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$; no → $-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

$IG = 0.778 - \frac{3}{5}(0.9181) - \frac{2}{5} = -0.1728$

finished HW: yes → $-\frac{2}{2}\log\frac{2}{2} - \frac{0}{2}\log\frac{0}{2} = 0$; no → $-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918$

$0.778 - 0.918\frac{3}{5} = 0.2272$

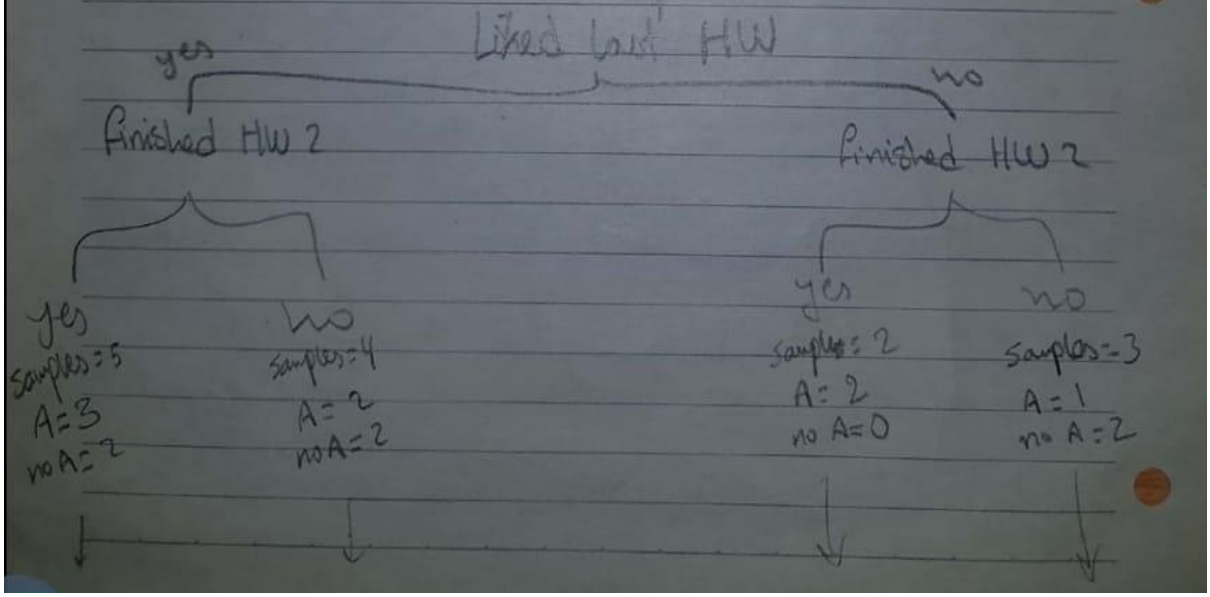
senior: yes → $-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$; no → $-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

$0.778 - \frac{3}{5} \times 0.918 - \frac{2}{5} = -0.1728$

likes coffee: yes → $-\frac{1}{1}\log 1 = 0$; no → $-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$

$0.778 - \frac{4}{5} = -0.022$

⇒ Finished HW has the highest IG



Scikit-Learn Tree:

- Split the dataset into X_train, X_test, y_train, y_test.
- Create a decision tree object from sklearn.tree.DecisionTreeClassifier.
- Use the decision tree object to fit the X_train and y_train.
- Use the test data to evaluate the model.
- To visualize the tree, set a max depth for the tree as the data is big and will cause problems while visualizing or will take a lot of time.

Decision Tree Built from Scratch:

Created class DecisionTree which takes in the data, target column, and the value of the output we're looking for (positive).

The class contains 3 functions: getEntropy, getGain, and updateTree.

- getEntropy calculates the entropy using the data given by getting the negative of the sum of the positive ration * log2 the positive ratio and the negative ratio * log2 the negative ratio.
- getGain calculates the information gain by subtracting the entropy of the data by the calculated mutual information.
- updateTree calls both functions and splits the tree according to the feature with the highest gain. Then, it recursively does the same for the children's trees.

To use:

Create object of the DecisionTree class given the needed info. and call the updateTree function.