Timothy Mah
8750634
txamah001@mymail.sim.edu.sg
CSIT375 Artificial Intelligence for Cybersecurity **Assignment Q3**

Steps for ID3

1. find info gain of each feature to find root
   - requires
     - base entropy
     - entropy of each feature

FORMULA FOR BASE ENTROPY

base entropy $= -\left\{[\text{probability of label 1} \times \log(\text{p of label 1})] + [\text{p of label 2} \times \log(\text{p label 2})] + \dots\right\}$

$= -\left\{\left[\frac{3}{12} \times \log\left(\frac{3}{12}\right)\right] + \left[\frac{8}{12} \times \log\left(\frac{8}{12}\right)\right] + \left[\frac{1}{12} \times \log\left(\frac{1}{12}\right)\right]\right\}$

('yes' label) ('unknown' label) ('no' label)

$= 1.38443$

entropy when file size = 100 $= -\left\{\left[\frac{1}{4} \times \log\left(\frac{1}{4}\right)\right] + \left[\frac{1}{4} \times \log\left(\frac{1}{4}\right)\right] + \left[\frac{2}{4} \times \log\left(\frac{2}{4}\right)\right]\right\}$ ('yes' label) ('unknown' label) ('no' label)

$= 1.5$

entropy when file size = 200 $= -\left\{\left[\frac{1}{4} \times \log\left(\frac{1}{4}\right)\right] + \left[\frac{2}{4} \times \log\left(\frac{2}{4}\right)\right]\right\}$ ('yes' label) ('no 'unknown' label) ('no' label)

$= 0.8112.7$

entropy when file size = 300 $= -\left\{\left[\frac{1}{4} \times \log\left(\frac{1}{4}\right)\right] + \left[\frac{1}{4} \times \log\left(\frac{1}{4}\right)\right] + \left[\frac{2}{4} \times \log\left(\frac{2}{4}\right)\right]\right\}$ ('yes' label) ('unknown' label) ('no' label)

$= 1.5$

entropy when file type = exe $= -\left\{\left[\frac{1}{7} \times \log\left(\frac{1}{7}\right)\right] + \left[\frac{1}{7} \times \log\left(\frac{1}{7}\right)\right] + \left[\frac{5}{7} \times \log\left(\frac{5}{7}\right)\right]\right\}$ ('yes' label) ('unknown' label) ('no' label)

$= 1.14883$

entropy when file = doc $= -\left\{\left[\frac{3}{8} \times \log\left(\frac{3}{8}\right)\right] + \left[\frac{1}{8} \times \log\left(\frac{1}{8}\right)\right] + \left[\frac{4}{8} \times \log\left(\frac{4}{8}\right)\right]\right\}$

$= 1.52192$

entropy when sections = 1 $= -\left\{\left[\frac{7}{8} \times \log\left(\frac{7}{8}\right)\right] + \left[\frac{1}{8} \times \log\left(\frac{1}{8}\right)\right]\right\}$

$= 0.91829$

entropy when sections = 3 $= \left\{\left[\frac{3}{3} \times \log\left(\frac{3}{3}\right)\right] + \left[\frac{0}{3} \times \log\left(\frac{0}{3}\right)\right]\right\}$

$= 1$

entropy when keyword = 5 $= \left\{\left[\frac{4}{7} \times \log\left(\frac{4}{7}\right)\right] + \left[\frac{2}{7} \times \log\left(\frac{2}{7}\right)\right] + \left[\frac{1}{7} \times \log\left(\frac{1}{7}\right)\right]\right\}$

$= 1.45914$

entropy when keyword = 2 $= -\left\{\left[\frac{1}{2} \times \log\left(\frac{1}{2}\right)\right]\right\}$

$= 0$

	Entropy	Malware = Yes	Malware = Unknown	Malware = No
H(C| File Size=100) x 4	1.5	1	1	2
H(C| File Size=300) x 4	1.5	1	1	2
H(C| File Size=200) X 4	0.81127	1	0	3
H(C| File Size)				
H(C| File Type=Executable) x7	1.14883	1	1	5
H(C| File Type=Document) x 5	1.52192	2	1	2
H(C| File Type)				
H(C| Number of Sections=1) x6	0.91829	0	2	4
H(C| Number of Sections=3) x6	1	3	0	3
H(C| Number of Sections)				
H(C| Suspicious Keyword=5) x6	1.45914	3	2	1
H(C| Suspicious Keyword=2) x6	0	0	0	6
H(C| Suspicious Keyword)				

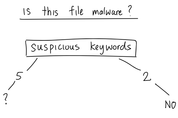
2. Build Tree

- sort features by entropy

Highest to lowest

File = doc	1.52192
Size = 100	1.5
Size = 300	1.5
SUS = 5	1.45914
File = exe	1.14883
Sections = 3	1
Sections = 1	0.91829
Size = 200	0.81127
SUS = 2	0

- lower entropy = higher info gain
- root has highest info gain
- calculate weighted entropy for each attribute, at each branch



entropy when SUS = 5 AND File = exe $= -\left\{\left[\frac{1}{2} \times \log\left(\frac{1}{2}\right)\right] + \left[\frac{1}{2} \times \log\left(\frac{1}{2}\right)\right]\right\}$

$= 1$

entropy when SUS = 5 AND file = doc $= -\left\{\left[\frac{1}{4} \times \log\left(\frac{1}{4}\right)\right] + \left[\frac{1}{4} \times \log\left(\frac{1}{4}\right)\right] + \left[\frac{1}{4} \times \log\left(\frac{1}{4}\right)\right]\right\}$

$= 1.5$

weighted entropy of SUS = 5 AND Filetype $= \left(\frac{6}{6} \times 1\right) + \left(\frac{1}{6} \times 1.5\right)$

$= 1.333$

entropy when SUS = 5 AND size = 100 $= -\left\{\left[\frac{1}{2} \times \log\left(\frac{1}{2}\right)\right] + \left[\frac{1}{2} \times \log\left(\frac{1}{2}\right)\right]\right\}$

$= 1$

entropy when SUS = 5 AND size = 300 $= -\left\{\left[\frac{1}{2} \times \log\left(\frac{1}{2}\right)\right] + \left[\frac{1}{2} \times \log\left(\frac{1}{2}\right)\right]\right\}$

$= 1$

entropy when SUS = 5 AND size = 200 $= -\left\{\left[\frac{1}{2} \times \log\left(\frac{1}{2}\right)\right] + \left[\frac{1}{2} \times \log\left(\frac{1}{2}\right)\right]\right\}$

$= 1$

weighted entropy of SUS = 5 AND File size $= \left(\frac{2}{6} \times 1\right) + \left(\frac{2}{6} \times 1\right) + \left(\frac{2}{6} \times 1\right)$

$= 1$

entropy when SUS = 5 AND sections = 3 $= -\left[\frac{3}{3} \times \log\left(\frac{3}{3}\right)\right]$

$= 0$

entropy when SUS = 5 AND sections = 1 $= -\left\{\left[\frac{1}{5} \times \log\left(\frac{1}{5}\right)\right] + \left[\frac{5}{5} \times \log\left(\frac{5}{5}\right)\right]\right\}$

$= 0.91829$

weighted entropy of SUS = 5 AND Number of Sections $= \left(\frac{3}{6} \times 0\right) + \left(\frac{3}{6} \times 0.91829\right)$

$= 0.45914$

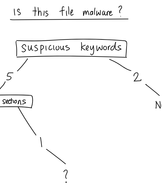
Sort by weighted entropy

file type = 1.33333

file size = 1

number of sections = 0.45914

Since number of sections has lowest weighted entropy, it has highest info gain



Based on intuition and deduction that the only unique element that uniquely identifies 'no' label when SUS keywords = 5 AND no. of sections = 1,

	File Size	File Type	Number of Sections	Suspicious Keyword	Malware
2	100	Document	1	5	Unknown
5	300	Executable	1	5	Unknown
8	200	Document	1	5	No

is file size = 200, hence next branch should be file size