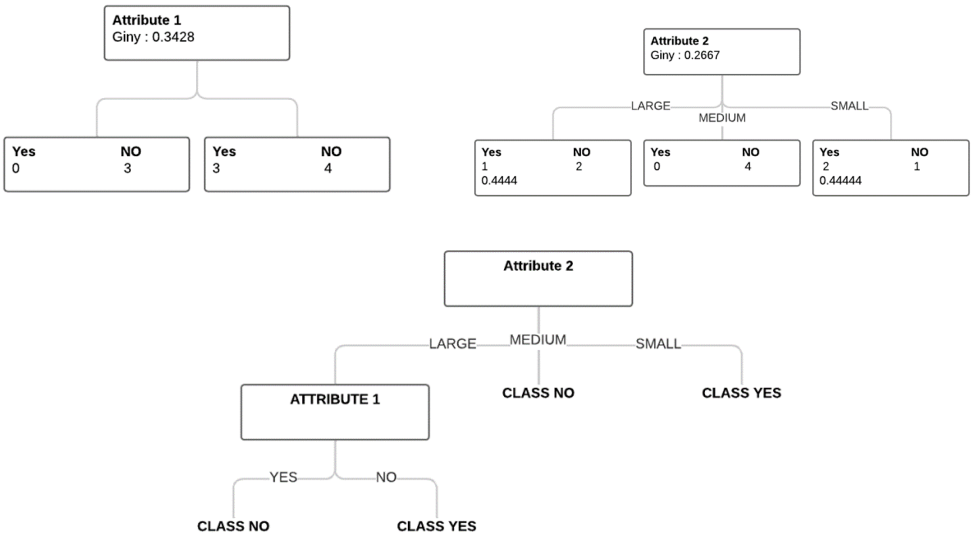
Machine Learning HW 10

DECISION TREES

Bui Phuong Thao - 11215341 - DSEB 63

1 Exercise 1

Build a decision tree based on given dataset using GINI IMPURITY



ATTRIBUTE 1	ATTRIBUTE 2	CLASS
NO	SMALL	YES
YES	MEDIUM	NO
YES	LARGE	NO
NO	SMALL	YES
NO	LARGE	YES

Build a decision tree based on given dataset using ENTROPHY

Entrophy(Class) = E(7,3) = $-(\frac{7}{10}\log_2\frac{7}{10}) - (\frac{3}{10}\log_2\frac{3}{10}) = 0.881$

Entrophy(Class, Attribute 1) = P(Yes).E(Yes) + P(No).E(No) = $\frac{3}{10}.E(0,3) + \frac{7}{10}.E(4,3) = 0 + 0.6896 = 0.6896$

Entrophy(Class, Attribute 2) = P(Large).E(Large) + P(Medium).E(Medium) + P(Small).E(Small) = $\frac{3}{10}.E(2,1) + \frac{4}{10}.E(4,0) + \frac{3}{10}.E(1,2) = 0.550$

Information gain for each split:

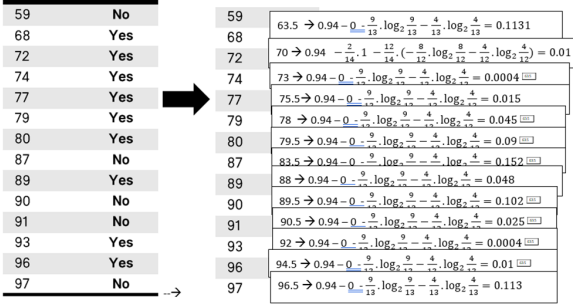
Gain(Class,Attribute1) = E(Class) - Entrophy(Class, Attribute 1) = 0.881 - 0.689 = 0.192

Gain(Class,Attribute2) = E(Class) - Entrophy(Class, Attribute 2) = 0.881 - 0.550 = 0.331

→ Choose Attribute 2 for the first split

2 Exercise 2

Humidity	Play Tennis?
59	No
68	Yes
72	Yes
74	Yes
77	Yes
79	Yes
80	Yes
87	No
89	Yes
90	No
91	No
93	Yes
96	Yes
97	No



Handling numerical attributes

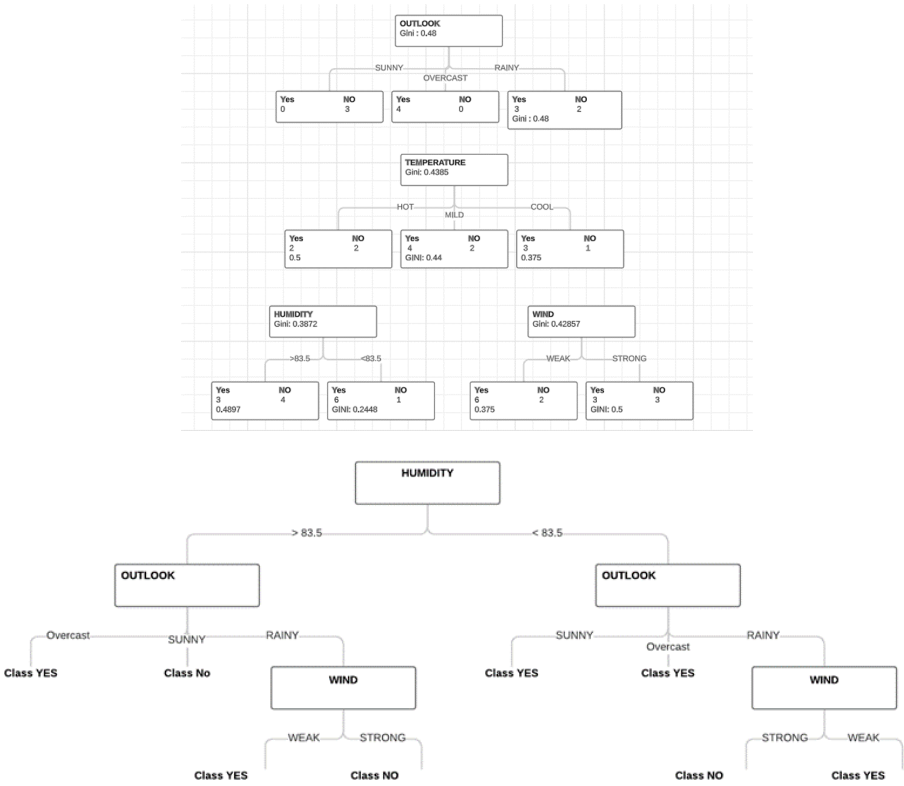
* sort the data in ascending order

* calculate the average of each adjacent pair of values

* compute the information gain with each splitting value to find the largest one

83.5 is the best splitting value with an information gain of 0.152, Humidity is now treated as a categorical attribute with two possible values

3 Exercise 3



Sample (Sunny, mild, 85, weak) belongs to class YES