SOWJANYA SADASHIVA
1001898874.

1a.

Data

Height	weight	age	gender.
1.6563	77.5687	37	M
1.5945	69.8713	33	W
1.6233	71.8122	33	m
1.6051	67.3119	26	W
1.6416	72.3383	30	m
1.5956	73.1325	27	W.

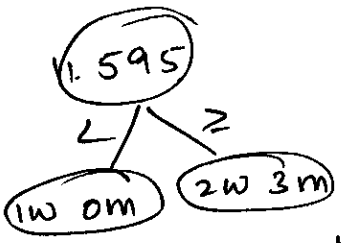
Entropy of entire data set = 1.
equal split.

a) Sort the "height" & calculate Ig.

height	gender.
1.5945	W
1.5956	W
1.6051	W
1.6233	M
1.6416	m
1.6563	m

→ i) calculate the mid value for (n-1).
i.e we get 5 values.
ii) calculate entropy
iii) calculate Ig.
iv) Pick the point with highest Ig for threshold.

$(1.5945 + 1.5956)/2 \approx 1.595$



cal entropy for each :- leaf node of this tree.

i.e $e_1 = -\frac{1}{1}\log_2\frac{1}{1} = 0$

$e_2 = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.9709$

weighted entropy $= \frac{1}{6}(0) + \frac{5}{6}(0.9709) =$

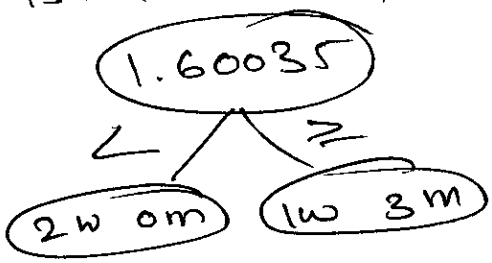
Ig : $E(S) - E(w)$

$= 1 - 0.809$

≈ 0.191

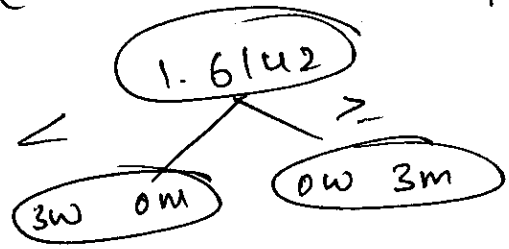
lly we calculate entropy for all data points in height attribute

b) $(1.5956 + 1.6051)/2 \approx 1.60035$



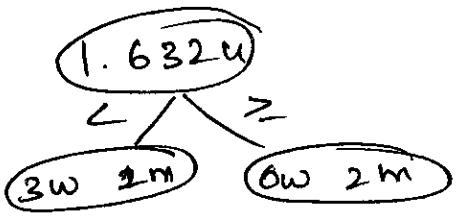
Ig = 0.45916

c) $(1.6051 + 1.6233)/2 \approx 1.6142$



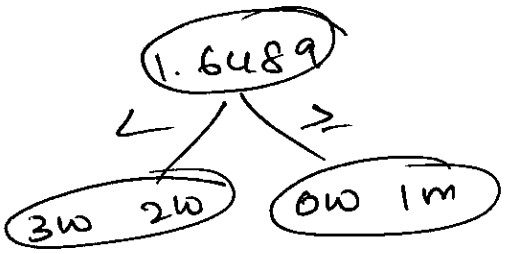
Ig = 1

d) $(1.6233 + 1.6416)/2 \approx 1.6324$



$Ig = 0.45916$

c) $(1.6416 + 1.6563)/2 \approx 1.6489$



$Ig = 0.1908$

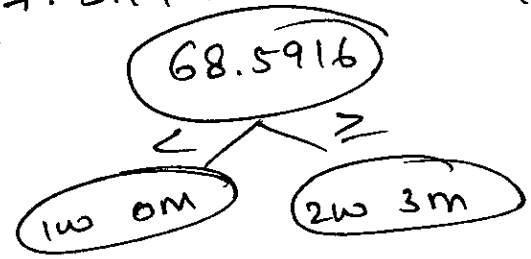
∴ from the calculation we know at this point

Height = Ig = 1 = 1.6142 (threshold).

Sort by "weight"

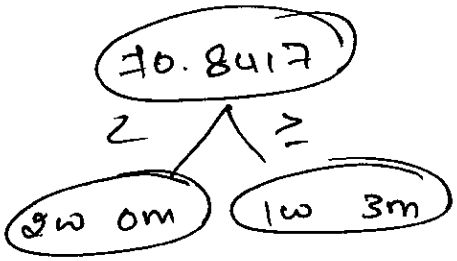
weight	gender
67.3119	w
69.8713	w
71.8122	M
72.3383	M
73.1325	w
77.5687	M

a) $(67.3119 + 69.8713)/2 \approx 68.5916$.



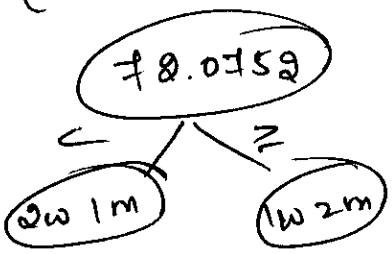
$Ig = 0.1908$

b) $(69.8713 + 71.8122)/2 \approx 70.8417$.



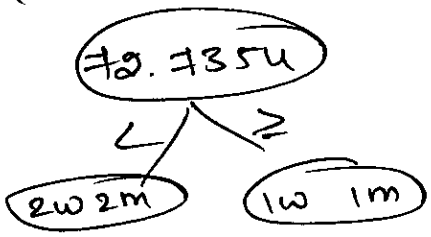
$Ig = 0.45916$

c) $(71.8122 + 72.3383)/2 \approx 72.0752$



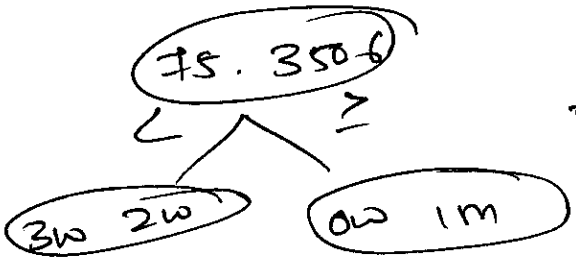
$Ig = 0.0817$

d) $(72.3383 + 73.1325)/2 \approx 72.7354$



$Ig = 0.0001$

e) $(73.1325 + 77.5687)/2 \approx 75.3506$



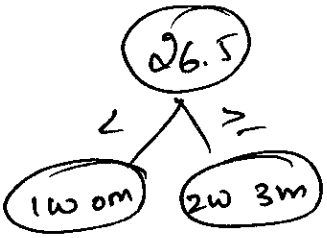
$Ig = 0.1908$

→ of the calculations. $Ig = 0.45916$ at 70.8417 is highest. we consider that as threshold.

Sort by "age"

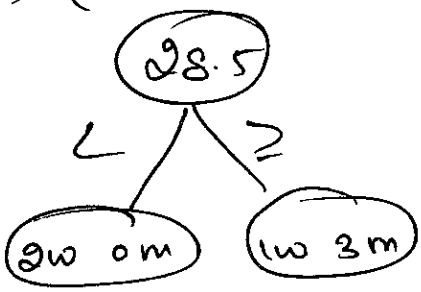
age	gender
26	W
27	W
30	M
33	W
33	M
37	M

a) $(26+27)/2 = 26.5$



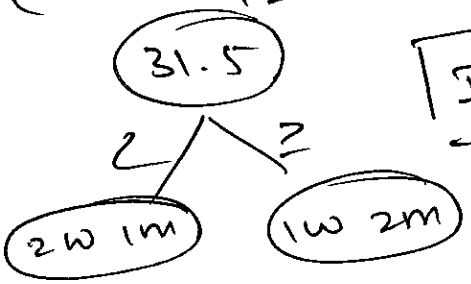
$Ig = 0.1908$

b) $(27+30)/2 \approx 28.5$



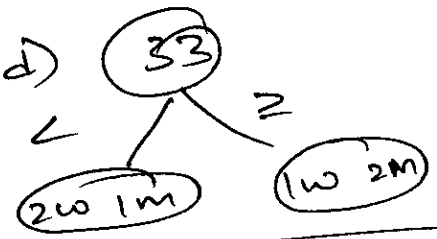
$Ig = 0.45916$

c) $(30+33)/2 = 31.5$



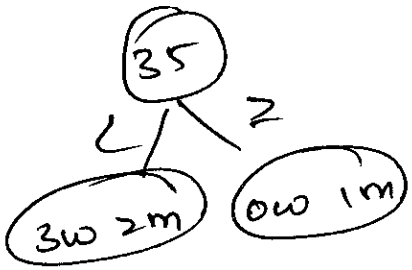
$Ig = 0.0817$

d)



$Ig = 0.0817$

e) $(33+37)/2 = 35$



$Ig = 0.1908$

→ 28.5 has highest $Ig = 0.45916$.

∴ we know

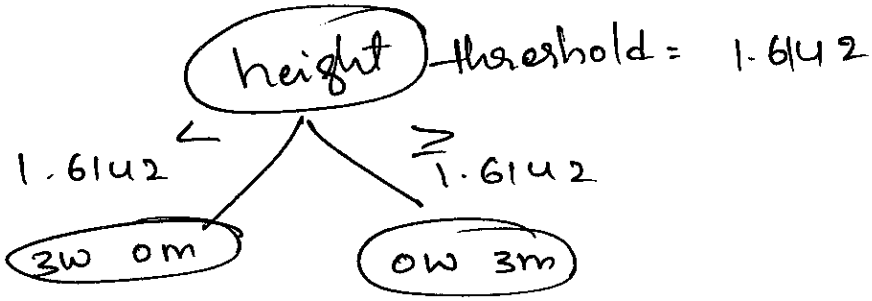
$Ig = [height] = 1 = 1.6142$

$Ig = [weight] = 0.45916 = \mp 0.8417$

$Ig = [age] = 0.45916 = 28.5$

of all height has highest.

∴ height = root node



* Since it is already a perfect split we need not continue.

* If the split is not perfect, we perform above steps repeated with the subsets we get for child-node till we have best purity at the leaf nod.