Question #1

A) employ 1R to answer the following
which attribute produces lowest error rate for predicting the outcome of >10 chinups?

First, build the table:

	yes	no		yes	no		yes	no	⋮	yes	no
weight			waist			heartrate			⋮	11	9
heavy	3/11	6/9	thin	9/11	2/9	fast	6/11	4/9	⋮		
light	8/11	3/9	wide	2/11	7/9	slow	5/11	5/9	⋮		

from the table, clearly waist has lowest error rate
rule #1: if waist = thin then >10 chinups = yes (9/11)
rule #2: if waist = wide then >10 chinups = no (7/9)

B) what is the error Rate of rules from (A)?
rule 1: 9/11 = 0.818, 1 - 0.818 = 0.182 or 18%.
rule 2: 7/9 = 0.778, 1 - 0.778 = 0.222 or 22%.

Question #2 Naive Bayes
A) What is probability >10 chinups = yes given weight = heavy, waist = wide, and heartrate = slow?
Bayes Theorem: $p(x|y) = \frac{p(y|x)\,p(x)}{p(y)}$

we want to know $p(>10\text{ chinups} = yes \mid heavy, wide, slow)$
$\Rightarrow p(yes \mid heavy, wide, slow) = \frac{p(heavy, wide, slow \mid yes)\,p(yes)}{p(heavy, wide, slow)}$

using marginal probability:
$p(heavy, wide, slow) = p(heavy, wide, slow \mid yes)p(yes) + p(heavy, wide, slow \mid no)p(no)$
$\Rightarrow p(yes \mid heavy, wide, slow) = \frac{p(heavy, wide, slow \mid yes)\,p(yes)}{p(heavy, wide, slow \mid yes)p(yes) + p(heavy, wide, slow \mid no)p(no)}$

assuming independence, $p(heavy, wide, slow \mid yes) = p(heavy \mid yes)\,p(wide \mid yes)\,p(slow \mid yes)$

$\Rightarrow p(yes \mid heavy, wide, slow) = \frac{3/11 \cdot \frac{2}{11} \cdot \frac{5}{11} \cdot \frac{11}{20}}{\left(\frac{3}{11} \cdot \frac{2}{11} \cdot \frac{5}{11} \cdot \frac{11}{20}\right) + \left(\frac{6}{9} \cdot \frac{7}{9} \cdot \frac{5}{9} \cdot \frac{9}{20}\right)} = 0.0373$

B) which values of weight, waist, and heartrate yield highest probability of >10 chinups = yes? answer: simply find the highest fractions for each attribute: light, thin, fast

#3 Use divide and conquer to construct a decision tree on the dataset, using Gini index as splitting criterion, to answer the following:

A) which attribute the best initial split? $Gini = 1 - \sum_i p_i^2$

weight: heavy: $\frac{3}{9}$ = yes, $\frac{6}{9}$ = no

light: $\frac{8}{11}$ = yes, $\frac{3}{11}$ = no

$Gini(weight) = 1 - \left(\frac{9}{20}\left(\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2\right) + \frac{11}{20}\left(\left(\frac{8}{11}\right)^2 + \left(\frac{3}{11}\right)^2\right)\right) = 0.4182$

waist: wide: $\frac{2}{9}$ = yes, $\frac{7}{9}$ = no

thin: $\frac{9}{11}$ = yes, $\frac{2}{11}$ = no

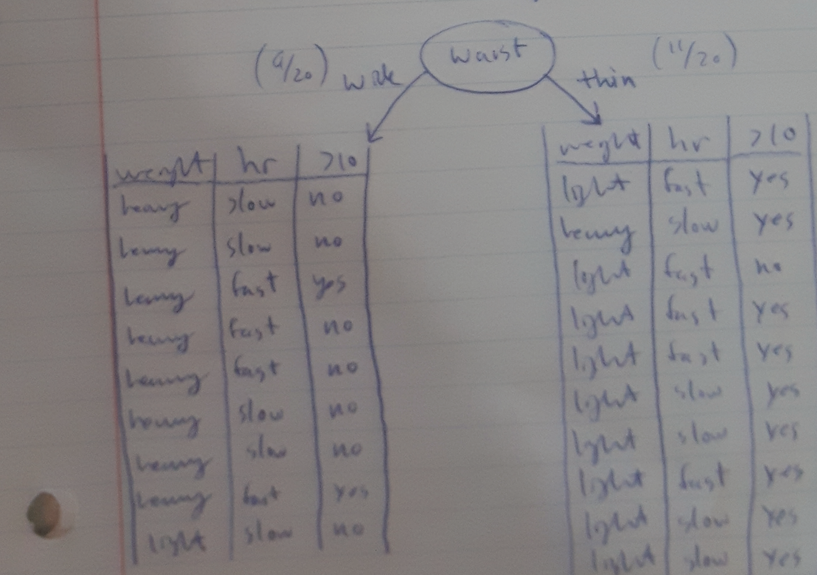
$Gini(waist) = 0.3192$

heartrate: fast: $\frac{6}{10}$ = yes, $\frac{4}{10}$ = no

slow: $\frac{5}{10}$ = yes, $\frac{5}{10}$ = no

$Gini(heartrate) = 0.49$

spliton waist as root.

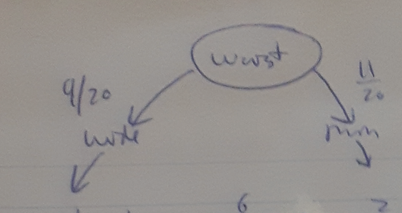


weight	hr	>10
heavy	slow	no
heavy	slow	no
heavy	fast	yes
heavy	fast	no
heavy	fast	no
heavy	slow	no
heavy	slow	no
heavy	fast	yes
light	slow	no

weight	hr	>10
light	fast	yes
heavy	slow	yes
light	fast	no
light	fast	yes
light	fast	yes
light	slow	yes
light	slow	yes
light	fast	yes
light	slow	yes
light	slow	yes
light	fast	no

because we split, we create 2 subsets of data based on instances that had waist = wide or waist = thin

We then calculate gini on these sub-sets to keep building tree



wide

weight: heavy: $\frac{6}{8}$ = no, $\frac{2}{8}$ = yes 8/9
light: $\frac{1}{1}$ = no, $\frac{0}{1}$ = yes 1/9

heartrate: slow: $\frac{5}{5}$ = no, $\frac{0}{5}$ = yes, 5/9
fast: 3/4 = no, 1/4 = yes, 4/9

$Gini(weight) = 1 - \left(\frac{8}{9}\left(\left(\frac{6}{8}\right)^2 + \left(\frac{2}{8}\right)^2\right) + \left(\frac{1}{9}\left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right)\right)\right) = 0.338$

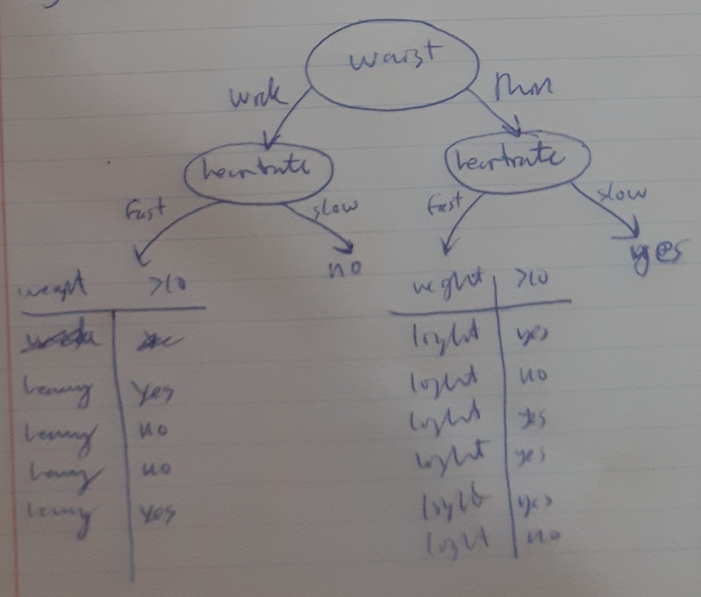
$Gini(heartrate) = 0.222 \Rightarrow$ heartrate

thin

weight: heavy 1/11, $\frac{1}{1}$ = yes, $\frac{0}{1}$ = no
light 10/11, $\frac{8}{10}$ = yes, $\frac{2}{10}$ = no
gini(weight) = 0.2909

$\Rightarrow$ heartrate

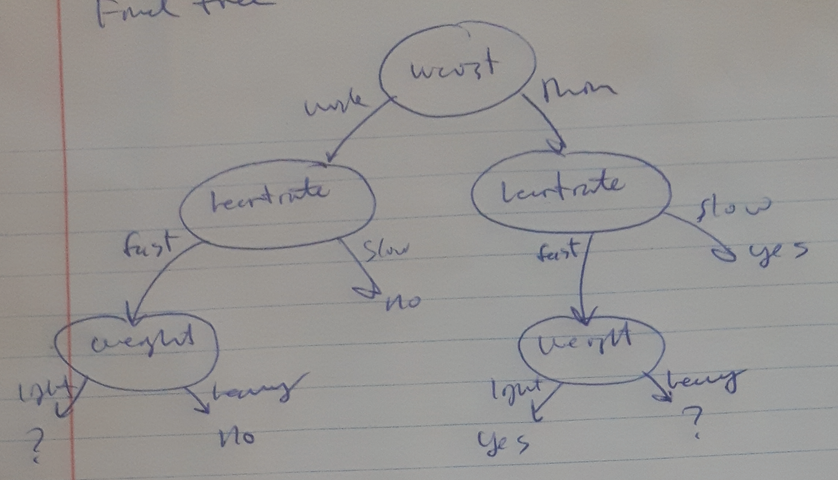
heartrate: fast 6/11, 4/6 = yes, 2/6 = no
slow 5/11, 5/5 = yes, 0/5 = no
gini(heartrate) = 0.242



weight	>10
~~wide~~	~~no~~
heavy	yes
heavy	no
heavy	no
heavy	yes

weight	>10
light	yes
light	no
light	yes
light	yes
light	yes
light	no

Final tree



Rules:
A → if waist = wide and heartrate = slow then >10 = no
B → if waist = thin and heartrate = slow then >10 = yes
C → if waist = wide and heartrate = fast and weight = heavy then >10 = no
D → if waist = thin and heartrate = fast and weight = light then >10 = yes

B) compare to 1R
rule #1 from 1R is a more general version of rule B from the tree

rule #2 from 1R is a more general version of rule A from the tree