DATASET

weight	**heartrate**	**>150 situps**
heavy	slow	*yes*
heavy	slow	*no*
heavy	fast	*no*
light	fast	*no*
heavy	slow	*yes*
heavy	fast	*no*
heavy	fast	*no*
light	fast	*no*
light	fast	*yes*
light	fast	*yes*
light	slow	*no*
light	slow	*yes*
light	fast	*yes*
heavy	slow	*no*
heavy	slow	*no*
heavy	fast	*yes*
light	slow	*no*
light	slow	*yes*
light	slow	*yes*
light	fast	*no*

consider the following dataset with 2 attributes (weight, heartrate) and two classes (>150 situps=yes, >150 situps=no)

Question 1
Using the 1R algorithm, answer the following:
a) build the table showing the errors for each attribute-value and each attribute overall
b) list the two rules given by the attribute with the lowest error rate from (a)

a)

attribute	yes	no	overall error
weight = heavy	3/9	6/9	3/20 + 5/20 = 8/20
weight=light	6/11	5/11	
heartrate=slow	5/10	5/10	5/20 + 4/20 = 9/20
heartrate=fast	4/10	6/10	

b) if weight=heavy then >150 situps=no
if weight=light then >150 situps=yes

Question 2 Naive Bayes
a) what is the probability >150 situps=yes given weight=heavy and heartrate=slow?
b) show

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

using the symmetry property:

$$P(A,B) = P(B,A)$$

and the product rule:

$$P(A,B) = P(A|B)P(B)$$

a)

$$p(yes|heavy, slow) = \frac{p(heavy|yes)p(slow|yes)p(yes)}{p(heavy|yes)p(slow|yes)p(yes)+p(heavy|no)p(slow|no)p(no)}$$

$$= \frac{\left(\frac{3}{9}\right)\left(\frac{5}{10}\right)\left(\frac{9}{20}\right)}{\left(\frac{3}{9}\right)\left(\frac{5}{10}\right)\left(\frac{9}{20}\right)+\left(\frac{6}{9}\right)\left(\frac{5}{10}\right)\left(\frac{11}{20}\right)} = 0.29$$

b) P(A|B)P(B) = P(B|A)P(A) => P(A|B) = P(B|A)P(A) / P(B)
(symmetry + product rule) (algebra)

Question #3
which attribute should be placed as the root for a decision tree of this dataset (using gini index)

solution) we need to check the gini index for both weight and heartrate
weight: weight=heavy (9 instances total, 3 yes ,6 no)
weight=light (11 instance total, 6 yes, 5 no)

=> gini(weight) =

$$1 - \frac{9}{20}\left(\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2\right) - \frac{11}{20}\left(\left(\frac{6}{11}\right)^2 + \left(\frac{5}{11}\right)^2\right) = 0.473$$

heartrate: heartrate=slow (10 instances,5 yes, 5 no)
heartrate=fast (10 instances, 4 yes, 6 no)

=> gini(heartrate) =

$$1 - \frac{10}{20}\left(\left(\frac{5}{10}\right)^2 + \left(\frac{5}{10}\right)^2\right) - \frac{10}{20}\left(\left(\frac{6}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right) = 0.49$$

=> weight should be placed as the root