weight	heartrate	>50 jumps?
heavy	slow	yes
heavy	slow	yes
heavy	fast	yes
light	fast	no
heavy	slow	yes
heavy	fast	no
heavy	fast	no
light	fast	no
light	fast	no
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

Midterm Exam Fall 2019 Wednesday October 9
CS405/505 Data Mining
Exam is worth 25% of final grade in course
Duration: 1 hour (60 minutes)
No mobile device or talking permitted during exam

Summary: the dataset provided has twenty instances (rows), two attributes (weight, heartrate) and two class outcomes (yes/no). Please answer the following questions:

Question #1 (A=25%, B=5%):

A) using the 1R algorithm, construct the table showing the rules predicting the outcome of >50 jumps for each attribute-value and the overall error rate for each attribute.

B) using the rules with lowest error from (A), predict the outcome of >50 jumps for following new instances:
instance 1: [weight=heavy, heartrate=slow]
instance 2: [weight=light, heartrate=fast]

Question #2 (A=20%, B=10%):

A) Using Naive Bayes algorithm, find P(>50 jumps=yes | weight=heavy, heartrate=slow)
in other words, predict the probability that >50 jumps=yes given the person is heavy and has slow heartrate

B) The denominator in Bayes' Theorem is the overall probability of the evidence, in 2A, this is the probability that the person is both heavy and slow:

$$P(yes|Evidence) = \frac{P(Evidence|yes)P(yes)}{P(Evidence)}$$

show, using the sum and product rules of probability, that:

$$P(Evidence) = P(Evidence|yes)P(yes) + P(Evidence|no)P(no)$$

hint - start with the sum rule

Question #3 (A=15%, B=10%):

A) Using the Gini index, find the attribute that should be placed as the root for a decision tree constructed on this dataset.

B) list the subset of instances that reach each child of the root node from (A)

Question #4 (A=10%, B=5%):

Consider the numerical dataset with two attributes (weight, heartrate) and numerical outcomes (number of jumps). please answer the following:

A) plot two separate scatter plots:
i) weight on x-axis, jumps on y-axis
ii) heartrate on x-axis, jumps on y-axis

B) by hand, estimate the best fit line for the two separate plots. which attribute has the higher correlation (r-value) with jumps?

weight	heartrate	jumps
191	50	60
189	52	60
193	58	101
162	62	37
189	46	58
182	56	42
211	56	38
167	60	40
176	74	40
154	56	250
169	50	38
166	52	115
154	64	105
247	50	50
193	46	31
202	62	120
176	54	25
157	52	80
156	54	73
138	68	43