Midterm

- Due Feb 21 at 10:30am
- Points 100
- Questions 15
- Available Feb 20 at 8:30am - Feb 21 at 10:30am
- Time Limit 75 Minutes
- Allowed Attempts 3

Instructions

Please read carefully before you start the test.

1. Multiple attempts are allowed only for extenuating circumstances such as power or internet loss, and I need to be notified when you have to make a second attempt.

2. You can use your notes, books, and slides.

3. The duration of the test is 75 minutes.

4. Please don't forget to submit your work after you finish the test. There's another link just below the Midterm for you to submit.

5. If you think a question is incorrect, please proceed with your assumption and notify me after the test. We can resolve those issues any time. So, don't let it slow you down during the test.

6. You can only see the answers when everyone is done with the test.

Attempt History

	Attempt	Time	Score
KEPT	Attempt 2	21 minutes	93 out of 100 *
LATEST	Attempt 2	21 minutes	93 out of 100 *
	Attempt 1	75 minutes	90 out of 100

* Some questions not yet graded

Score for this attempt: 93 out of 100 *
* Some questions not yet graded
Submitted Feb 21 at 9:35am
This attempt took 21 minutes.

Question 1

5 / 5 pts

Which of the following kinds of data can be mined?

○ Time Series

○ None of the above

○ Web

Correct!

● All of the anove

○ Text

Question 2

5 / 5 pts

Please match the terms which have the same meaning?

Correct!

Object

Instance

Correct!

Attribute

Feature

Other Incorrect Match Options:

- size
- correlation

Question 3

10 / 10 pts

	Play Basketball	Doesn't Play Basketball	Sum(row)
Cereal	213	203	**416**
Not cereal	138	110	**248**
Sum(col.)	**351**	**313**	**664**

According to the table above, please find the expected number of students who do not eat cereal, but play basketball. Please round your result to the nearest integer.

○ 117

○ 50

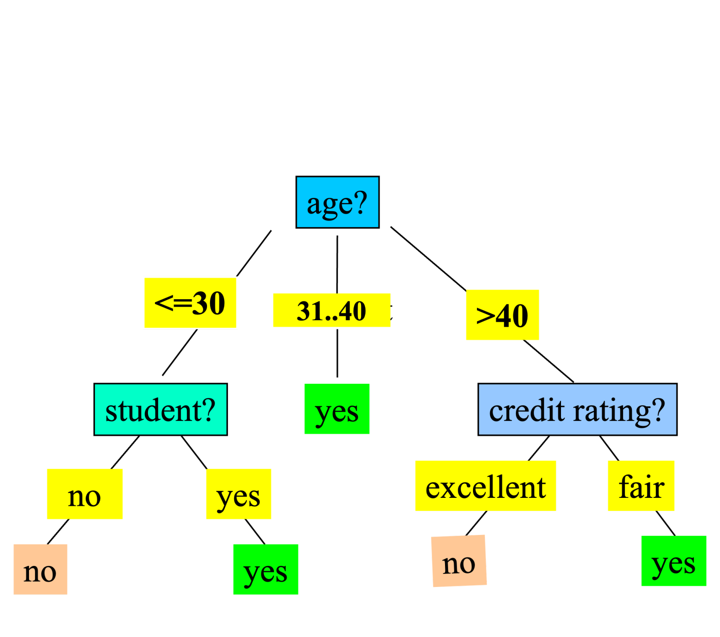
○ 41

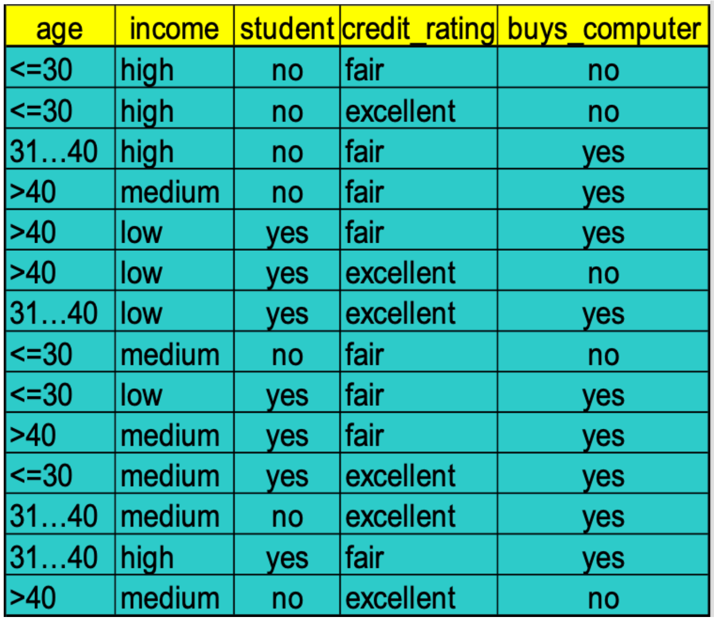
Correct!

● 131

Question 4

10 / 10 pts





age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Suppose you have an individual who is a student over 40 years old with low income level and fair credit rating. What is the predicted class by the decision tree above for this instance? Then, determine whether it is true positive (TP), false positive (FP), true negative (TN), or false negative (TN) when you check the row for this specific individual in the table right. (Please assume that yes=positive, no=negative)

○ no, FN

○ yes, FP

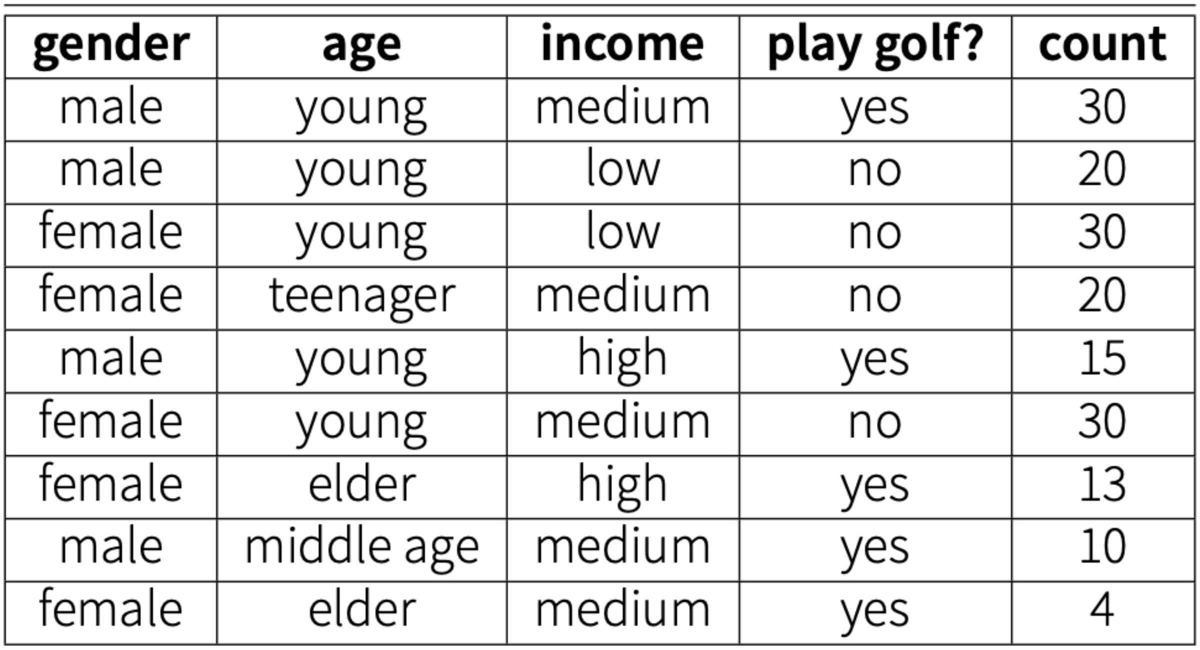
Correct!

● yes, TP

○ no, TN

Question 5

10 / 10 pts



gender	age	income	play golf?	count
male	young	medium	yes	30
male	young	low	no	20
female	young	low	no	30
female	teenager	medium	no	20
male	young	high	yes	15
female	young	medium	no	30
female	elder	high	yes	13
male	middle age	medium	yes	10
female	elder	medium	yes	4

In the table above, gender, age, and income are three attributes. Playing golf is the class label indicating that whether someone plays golf or not based on the attributes we have. The count column indicates the number of individuals who fall into their corresponding combinations. For instance, first row represents that there are 30 young males with medium income who play golf. Then, what is Gini(D) for the data set, D, presented in this table?

○ $-\frac{45}{172} \cdot \log(\frac{45}{172}) - \frac{72}{172} \cdot \log(\frac{72}{172})$

Correct!

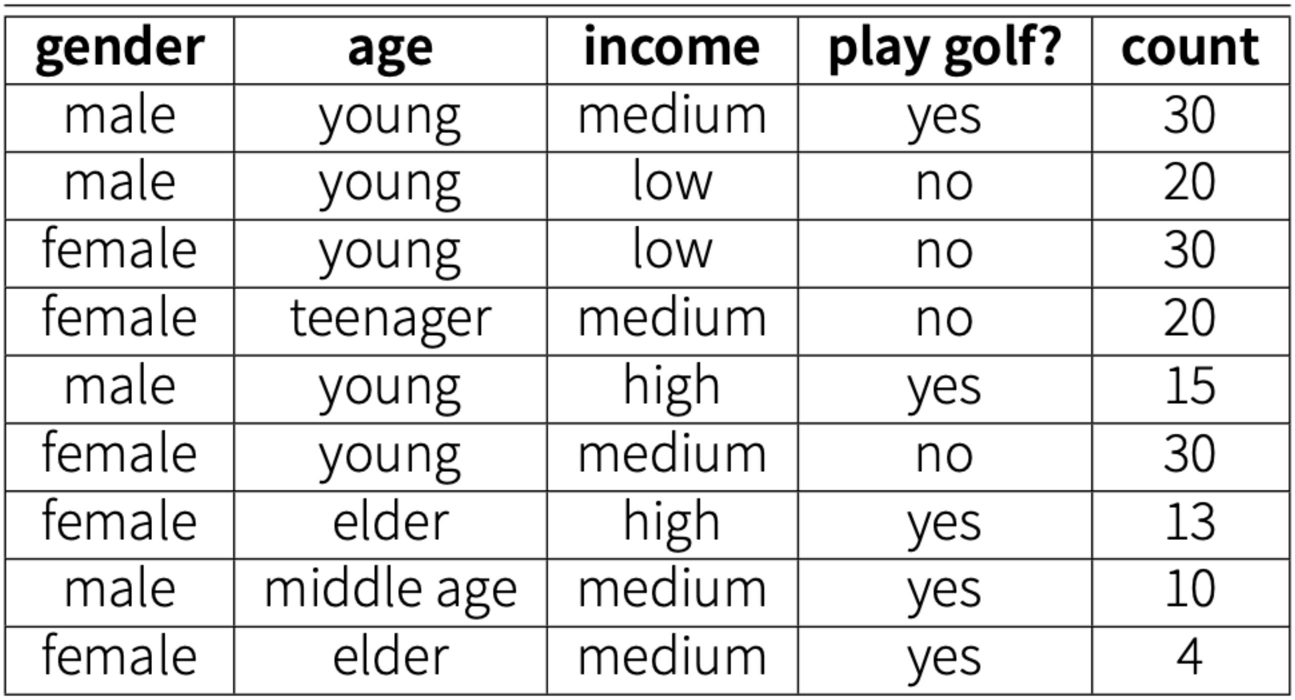
● $1 - (\frac{100}{172})^2 - (\frac{72}{172})^2$

○ $1 - (\frac{30}{72})^2 - (\frac{42}{72})^2$

○ $-\frac{50}{112} \cdot \log(\frac{50}{112}) - \frac{70}{112} \cdot \log(\frac{70}{112})$

Question 6

10 / 10 pts



gender	age	income	play golf?	count
male	young	medium	yes	30
male	young	low	no	20
female	young	low	no	30
female	teenager	medium	no	20
male	young	high	yes	15
female	young	medium	no	30
female	elder	high	yes	13
male	middle age	medium	yes	10
female	elder	medium	yes	4

In the table above, gender, age, and income are three attributes. Playing golf is the class label indicating that whether someone plays golf or not based on the attributes we have. The count column indicates the number of individuals who fall into their corresponding combinations. For instance, first row represents that there are 30 young males with medium income who play golf. Then, which of the followings would be enough to calculate information for income ($\text{info}_{\text{income}}$).

Correct!

● $\frac{94}{172} \cdot (-\frac{44}{94} \cdot \log(\frac{44}{94}) - \frac{50}{94} \cdot \log(\frac{50}{94}))$

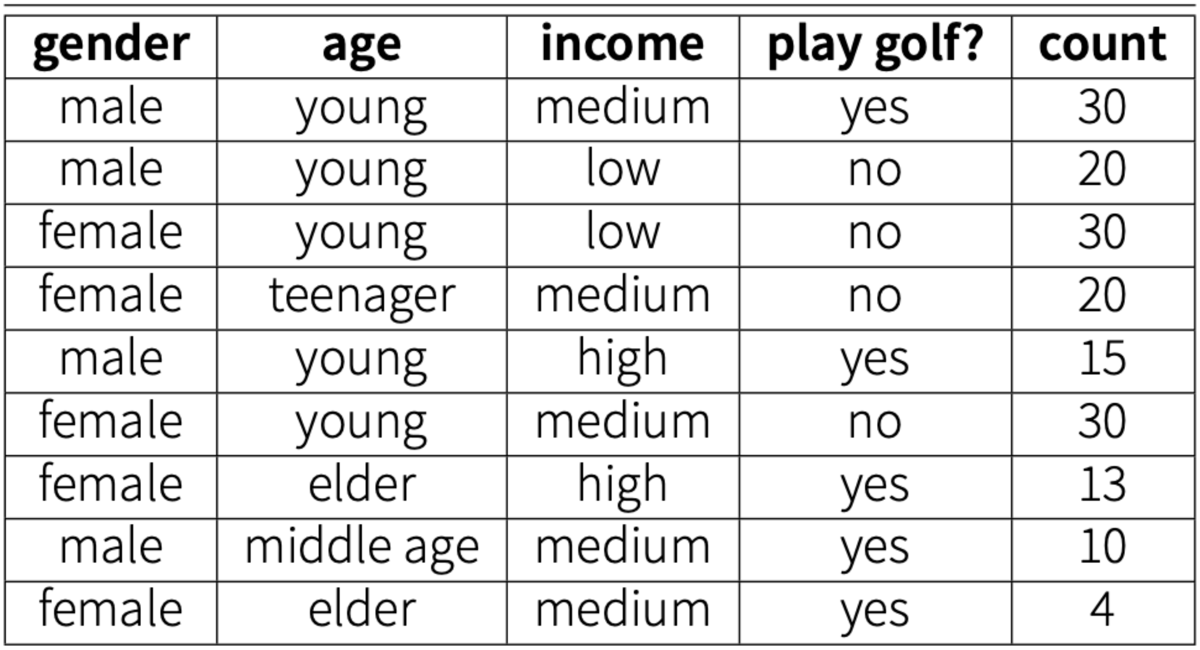
○ $\frac{50}{172} \cdot (-\frac{50}{50} \cdot \log(\frac{50}{50}) - \frac{0}{50} \cdot \log(\frac{0}{50}))$

○ $\frac{94}{172} \cdot (-\frac{28}{94} \cdot \log(\frac{28}{94}) - \frac{44}{94} \cdot \log(\frac{44}{94}))$

○ $\frac{28}{172} \cdot (-\frac{28}{28} \cdot \log(\frac{28}{28}) - \frac{0}{28} \cdot \log(\frac{0}{28}))$

Question 7

10 / 10 pts



gender	age	income	play golf?	count
male	young	medium	yes	30
male	young	low	no	20
female	young	low	no	30
female	teenager	medium	no	20
male	young	high	yes	15
female	young	medium	no	30
female	elder	high	yes	13
male	middle age	medium	yes	10
female	elder	medium	yes	4

In the table above, gender, age, and income are three attributes. Playing golf is the class label indicating that whether someone plays golf or not based on the attributes we have. The count column indicates the number of individuals who fall into their corresponding combinations. For instance, first row represents that there are 30 young males with medium income who play golf. Then, which is the Info(D)?

○ $-\frac{50}{112} \cdot \log(\frac{50}{112}) - \frac{70}{112} \cdot \log(\frac{70}{112})$

○ $-\frac{30}{72} \cdot \log(\frac{30}{72}) - \frac{42}{72} \cdot \log(\frac{42}{72})$

Correct!

● $-\frac{100}{172} \cdot \log(\frac{100}{172}) - \frac{72}{172} \cdot \log(\frac{72}{172})$

○ $-\frac{45}{172} \cdot \log(\frac{45}{172}) - \frac{72}{172} \cdot \log(\frac{72}{172})$

Question 8

10 / 10 pts

Suppose you have a data column which has its lowest value 20 and greatest value 100. If you apply min-max normalization to map all values to [0,1], what would be the normalized value of 20?

○ 0.8

Correct!

● 0

○ 0.2

○ 0.5

Question 9

10 / 10 pts

	Test 1	Test 2	Test 3	Test 4	Test 5
Adam	1	0	1	0	1
Eve	1	0	1	1	0
Mary	1	0	0	0	1

In the table above all tests are symmetric binary attributes for three patients. What is the distance (dissimilarity) between Mary and Eve?

Correct!

● 0.6

○ 0.1

○ 0.4

○ 0.2

Question 10

8 / 8 pts

Suppose you have 29, 75, 12, 20, 168, 163, 140, 52, 4, 37, 36, 123, 120, 31, 111. Then perform smoothing by bin boundaries with a bin depth of 3. Which is the smoothed second bin?

Correct!

● 29, 29, 36

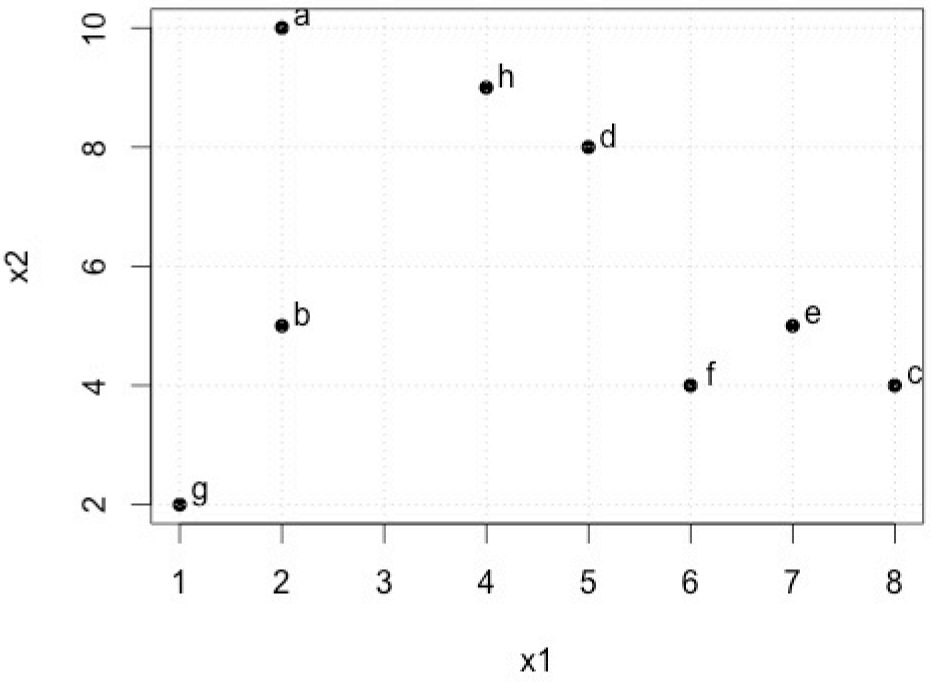
○ 31,31,31,31

○ 31, 31, 36

○ 31,31,31

Question 11

0 / 5 pts



What is the Manhattan distance between points f and c?

○ 8

○ 6

Correct Answer

○ 2

You Answered

● 4

Question 12

3 / 3 pts

Suppose we have the ordinal values, tall, grande, venti, and trenta, in increasing order for the size of coffee cups. If you map each size to a numerical value, which of the followings corresponds to grande?

○ 0.66

○ 0

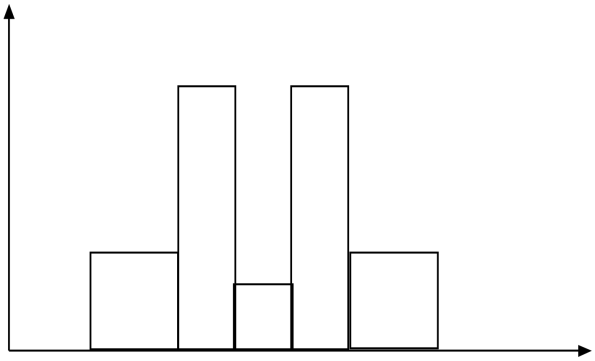
○ 1

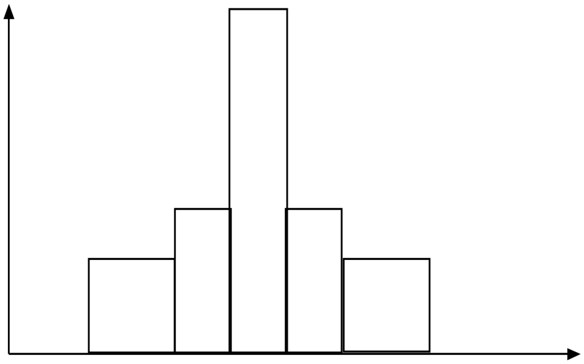
Correct!

● 0.33

Question 13

2 / 2 pts





Two histograms above belong to two different variables. Then, it's not possible for these two variables to have the same box plot representation.

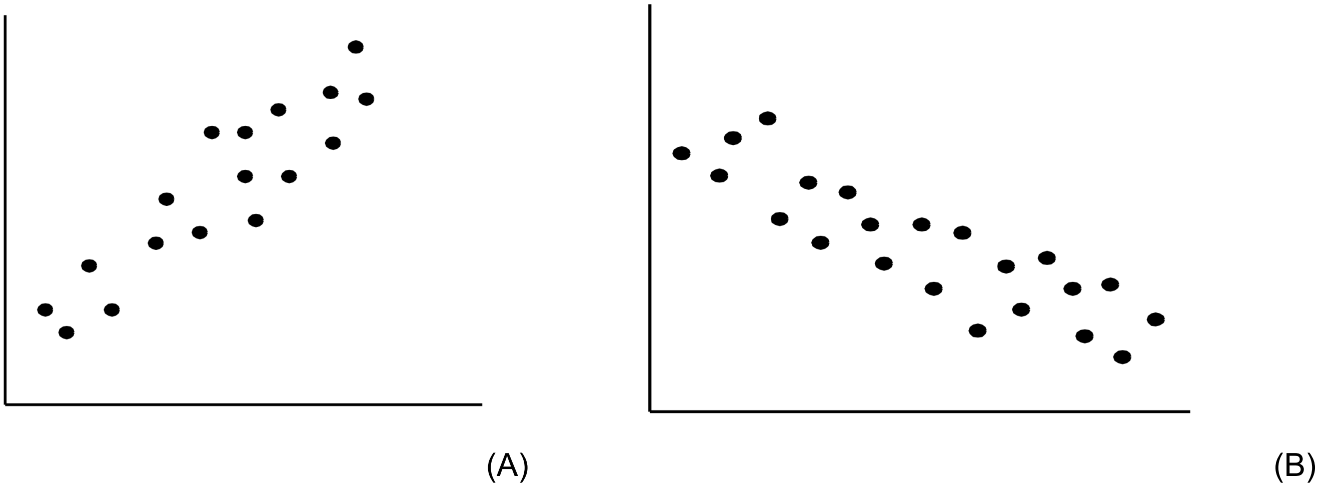
○ True

Correct!

● False

Question 14

Not yet graded / 2 pts



Please comment on differences on the scatter plots, A and B, above in terms of correlations.

Your Answer:

B is positive correlation

A is negative correlation

Question 15

0 / 0 pts

Do you remember that you need to upload your work after finishing the test?

○ no

Correct!

● yes

Quiz Score: 93 out of 100

* Some questions not yet graded