Question **1**

Which of the following steps is NOT how to create or use random forests?

Select one:
a. Select a portion of the features at random with replacement for the training
b. Generate complete trees without pruning
c. Select a portion of examples at random with replacement for the training
d. When using the random forest for classification, the final output is the most popular class
e. Decide on the number of features using a validation set

The correct answer is: Select a portion of the features at random with replacement for the training

Question **2**

Suppose we accidentally swap the columns with the rows when generating a confusion matrix, then all the correct cases are still the diagonals.

Select one:
True
False

The correct answer is 'True'.

Question **3**

The range of possible values for AUC is from -1 to 1.

Select one:
True
False

The correct answer is 'False'.

Question **4**

For any classifier, its confusion matrix is always a 2 by 2 matrix.

Select one:
True
False

The correct answer is 'False'.

Question **5**

Term frequency (TF) is document specific.

Select one:
True
False

The correct answer is 'True'.

Question **6**

Given Bayes' Rule:

$$p(C = c \mid \mathbf{E}) = \frac{p(\mathbf{E} \mid C = c) \cdot p(C = c)}{p(\mathbf{E})}$$

What is the term best used to describe p(C=c)?

Select one:
a. Conditional probability
b. Righthand probability
c. Conditional independence probability
d. Prior probability
e. Posterior probability

The correct answer is: Prior probability

Question **7**

With naïve Bayes, we assume that the predictor variables are conditionally independent of one another given the response value.

Select one:
True
False

The correct answer is 'True'.

Question **8**

What is the result of this R expression?

sum(c(TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, FALSE, FALSE, TRUE))

The correct answer is: 6

Question **9**

How many trigrams are in the sentence: "Lehigh beat Lafayette last year"?

The correct answer is: 3

Question **10**

A new web search engine called “Lehigh-Finds” is at a competitive disadvantage against Google and Bing because Lehigh-Finds lacks the click-through data to learn models of what answers are better than others.

Select one:
True
False

The correct answer is 'True'.

Question **11**

The point of analytical engineering is to develop complex solutions by addressing every possible contingency.

Select one:
True
False

The correct answer is 'False'.

Question **12**

Which of the following is the correct expected value formula for targeted marketing?

Select one:

a. Expected value = pR(x)·vR + (1-pR(x))·vNR

b. Expected value = pR(x) + vR·(1+pR(x)) + vNR

c. Expected value = pR(x)·vR + (1+pR(x))·vNR

d. Expected value = pR(x) + vR ·(1-pR(x))·vNR

The correct answer is: Expected value = pR(x)·vR + (1-pR(x))·vNR

Question **13**

Using N-grams greatly decreases the size of the feature set.

Select one:

True
False

The correct answer is 'False'.

Question **14**

Selenium enables R to extract information from websites and can even put data into website forms.

Select one:
True
False

The correct answer is 'True'.

Question **15**

The point of analytical engineering is to:Select one:

a. None of the others are appropriate answers
b. Promote thinking about problems data analytically
c. Develop complex solutions by addressing every possible contingency
d. Analyze engineers

The correct answer is:
Promote thinking about problems data analytically

Question **16**

We saw that n-grams can be useful when predicting who wrote a document.

Select one:
True
False

The correct answer is 'True'.

Question **17**

Churn causes a problem because it causes companies to lose money.

Select one:
True
False

The correct answer is 'True'.

Question **18**

A shapefile contains one or more series of points that, when connected, form the outline of a polygon.

Select one:
True
False

The correct answer is 'True'.

Question **19**

ROC graphs have the particular advantage that anyone can interpret them with ease.

Select one:
True
False

The correct answer is 'False'.

Question **20**

The y-axis of an ROC curve is theSelect one:

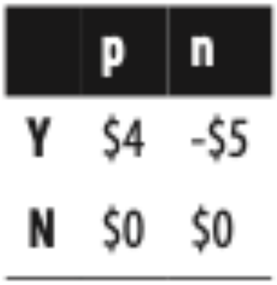
a. True negative rate

b. False positive rate
c. None of the others contain the correct answer
d. False negative rate
e. True positive rate

The correct answer is: True positive rate

Question **21**

Consider the situation in which you are targeting donors from recent college graduates, but you've chosen the high touch approach using phone calls and a professional telemarketing staff. Each person the telemarketer reaches incurs a charge of $5 (for the telemarketer), and when a donation occurs it is fairly small, just $9 on average. Thus, the cost/benefit matrix is:



	p	n
Y	$4	-$5
N	$0	$0

Imagine now the profit curve that we would draw given a classifier to target the population that the telemarketer calls. True or False: The profit curve could include points that showed a loss of $1,000 or more.

Select one:
True
False

The correct answer is 'True'.