DS-200 Q&As

Data Science Essentials

Pass Cloudera DS-200 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.pass2lead.com/ds-200.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Cloudera Official Exam Center

QUESTION 1

You have a large file of N records (one per line), and want to randomly sample 10% them. You have two

functions that are perfect random number generators (through they are a bit slow):

Random_uniform () generates a uniformly distributed number in the interval [0, 1] random_permotation (M)

generates a random permutation of the number O through M -1.

Below are three different functions that implement the sampling.

Method A

```
For line in file: If random_uniform ()
```

Method B

```
i = 0
for line in file:
    if i % 10 = = 0;
        print line
    i += 1
```

Method C

```
idxs = random_permotation (N) [: (N/10)]
i = 0
for line in file:
    if i in idxs:
        print line
    i +=1
```

Which method might introduce unexpected correlations?

A. Method A

B. Method B

C. Method C

Correct Answer: C

QUESTION 2

You have a large file of N records (one per line), and want to randomly sample 10% them. You have two functions that are perfect random number generators (through they are a bit slow):

Random_uniform () generates a uniformly distributed number in the interval [0, 1] random_permotation (M) generates a random permutation of the number O through M -1.

Below are three different functions that implement the sampling.

Method A

```
For line in file: If random_uniform ()
```

Method B

```
i = 0
for line in file:
if i % 10 = = 0;
print line
i += 1
```

Method C

```
idxs = random_permotation (N) [: (N/10)]
i = 0
for line in file:
if i in idxs:
print line
i +=1
```

Which method is least likely to give you exactly 10% of your data?

A. Method A

B. Method B

C. Method C

Correct Answer: B

QUESTION 3

What are three benefits of running feature selection analysis before filtering a classification model?

A. Eliminates the need to include a regularization term

B. Reduces the number of subjective decisions required to construct the model

C. Guarantee the optimally of the final model

D. Speeds up the model fitting process

E. Develops an understanding of the importance of different features

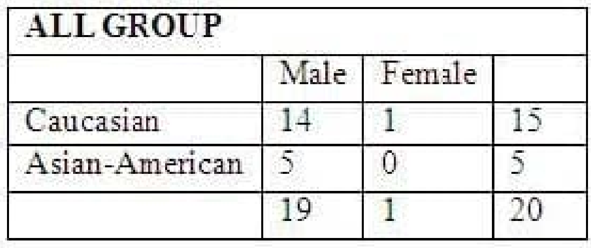
F. Improves the predictive performance of the model

Correct Answer: DEF

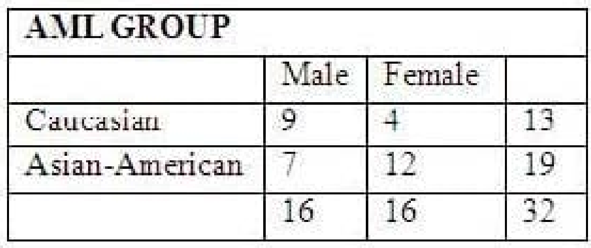
QUESTION 4

There are 20 patients with acute lymphoblastic leukemia (ALL) and 32 patients with acute myeloid leukemia (AML), both variants of a blood cancer.

The makeup of the groups as follows:



ALL GROUP			
	Male	Female	
Caucasian	14	1	15
Asian-American	5	0	5
	19	1	20



AML GROUP			
	Male	Female	
Caucasian	9	4	13
Asian-American	7	12	19
	16	16	32

Each individual has an expression value for each of 10000 different genes. The expression value for each gene is a continuous value between -1 and 1.

With which type of plot can you encode the most amount of the data visually?

You choose to perform agglomerative hierarchical clustering on the 10,000 features. How much RAM do you need to hold the distance Matrix, assuming each distance value is 64-bit double?

A. ~ 800 MB

B. ~ 400 MB

C. ~ 160 KB

D. ~ 4 MB

Correct Answer: B

QUESTION 5

You have just run a MapReduce job to filter user messages to only those of a selected geographical region. The output for this job in a directory named westUsers, located just below your home directory in HDFS. Which command gathers these records into a single file on your local file system?

A. Hadoop fs getmerge westUsers WestUsers.txt

B. Hadoop fs get westUsers WestUsers.txt

C. Hadoop fs cp westUsers/* westUsers.txt

D. Hadoop fs getmerge R westUsers westUsers.txt

Correct Answer: B