4 BIG DATA PROJECTS

1. An executive describes her company's "low latency, multiple terabyte" requirements for managing Big Data. To which characteristics of Big Data is the executive referring?
   (A) Velocity and variety.
   (B) Volume and variety.
   (C) Volume and velocity.

2. When evaluating the fit of a machine learning algorithm, it is most accurate to state that:
   (A) accuracy is the ratio of correctly predicted positive classes to all predicted positive classes.
   (B) recall is the ratio of correctly predicted positive classes to all actual positive classes.
   (C) precision is the percentage of correctly predicted classes out of total predictions.

3. Karlsson also learns of the model measure of accuracy. Based on Exhibit 1, Karlsson's model's accuracy metric is *closest* to:
   (A) 79%.
   (B) 89%
   (C) 69%.

4. Which of the following uses of data is most accurately described as curation'?
   (A) A data technician accesses an offsite archive to retrieve data that has been stored there
   (B) An investor creates a word cloud from financial analysts' recent research reports about a company.
   (C) An analyst adjusts daily stock index data from two countries for their different market holidays.

   Freja Karlsson is a bond analyst with Storbank AB. Over the past several months, Karlsson has been working to develop her own machine learning (ML) model that she plans to use to predict default of the various bonds that she

covers. The inputs to the model are various pieces of financial data that Karlsson has compiled from multiple sources.

After Karlsson has constructed the model using her knowledge of appropriate variables, Karlsson runs the model on the training set. Each firm's bonds are classified as predicted-to-default or predicted-not-to-default. When Karlsson's model predicts that a bond will default and the bond actually defaults, Karlsson considers this to be a true positive. Karlsson then evaluates the performance of her model using error analysis. The confusion matrix that results is shown in Exhibit 1.

N = 474		Actual bond Status	
		Bond Default	No Default
Model Prediction	Bond Default	307	31
	No Default	23	113

5. Under which of these conditions is a machine learning model said to be underfit?
   (A) The model identifies spurious relationships.
   (B) The input data are not labelled.
   (C) The model treats true parameters as noise.

6. Based on Exhibit 1, Karlsson's model's precision is *closest* to:
   (A) 71%
   (B) 81%
   (C) 91%

7. In big data projects, data exploration is least likely to encompass:
   (A) feature design.
   (B) feature engineering.
   (C) feature selection.

8. Karlsson is especially concerned about the possibility that her model may indicate that a bond will not default, but then the bond actually defaults. Karlsson decides to use the model's recall to evaluate this possibility. Based on the data in Exhibit 1, the model's recall is closest to:
   (A) 83%
   (B) 73%
   (C) 93%

9. The process of splitting a given text into separate words is best characterized as:

(A) tokenization.
(B) bag-of-words.
(C) stemming.

10. Karlsson would like to gain a sense of her model's overall performance. In her research, Karlsson learns about the F1 score, which she hopes will provide a useful measure. Based on Exhibit 1, Karlsson's model's F1 score is closest to:
(A) 72%
(B) 82%
(C) 92%

11. Big data is most likely to suffer from low:
(A) variety.
(B) velocity.
(C) veracity.