4 BIG DATA PROJECTS

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

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. The process of splitting a given text into separate words is best characterized as:
   (A) tokenization.
   (B) bag-of-words.
   (C) stemming.

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.

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

6. 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%

7. 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%

8. 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%.

9. 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.

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

11. 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.