DLI Accelerated Data Science Teaching Kit

Module 3 Quiz:
Data Pre-processing (ETL)

1. Why do we need data pre-processing?

A. Data lacks attributes or contains missing values
B. Data contains incorrect records such as outliers
C. Data contains conflicting records or discrepancies.
D. Data has thousands of attributes.

Answer: ABC

Explanation: See lecture 3.1 slides

2. What are the basic steps for data preprocessing?

A. Data cleaning and/or statistical preprocessing
B. Feature selection
C. Feature reduction
D. Building machine learning models on the data

Answer: ABC

Explanation: See lecture 3.1 slides

3. What are challenges introduced by missing values?

A. Missing values are generated by collection errors or missing observations.
B. Missing values are the most common problem in data analytics.
C. When missing values come out, certain model performance is decreased.
D. Even for models that can handle missing values, they might be sensitive to it.

Answer: ABCD

Explanation: See lecture 3.2 slides

4. Which method can be used to process missing values?

A. Dummy substitution
B. Mean substitution
C. Supervised learning
D. Frequent substitution

Answer: ABD

Explanation: See lecture 3.2 slides

5. What are the benefits of feature selection?

A. Simplifying the models
B. Shortening time for model construction
C. Avoiding curse of dimensionality
D. Enhancing model generalization

Answer: ABCD

Explanation: See lecture 3.4 slides

6. Which feature reduction methods are unsupervised?

A. Principal Component Analysis (PCA)
B. Independent Component Analysis (ICA)
C. Autoencoder
D. Linear Discriminant Analysis (LDA)

Answer: ABC

Explanation: See lecture 3.6 slides

7. Data dirtiness can affect the overall results when querying large data sets

A. True
B. False

Answer: A

8. Data Wrangler is a scripting language used for data cleaning and transformation.

A. True
B. False

Answer: B