Subject number : XXXXXXXXX

Sujet corrigé - 1 -Introduction to Data

This is the **question paper**. It is **NOT** the answer sheet.
Please check that the number on your question paper matches the number on your answer sheet. To complete the answer sheet correctly, you must:

- use a **black** ink pen
- shade in the boxes **completely without going over the edges**
- if you make a mistake, erase the box with a whiteout ("Tipp-Ex"), **but do not redraw it**
- every question has a **unique correct answer**.

Box correctly ticked

Box incorrectly ticked

1 **Which of the following is an example of structured data?**
1 Point - Only one correct choice

- ☐ **A.** Free-text doctor notes
- ☒ **B.** Lab test results in a spreadsheet
- ☐ **C.** GPS traces from a running app
- ☐ **D.** Social media posts

> Structured data are organized in a strict schema (rows/columns), e.g., spreadsheets.

2 **What is semi-structured data?**
1 Point - Only one correct choice

- ☐ **A.** Data with a rigid schema only
- ☐ **B.** Data with no recognizable pattern
- ☒ **C.** Data with some organizational tags but flexible fields
- ☐ **D.** Data only in images or videos

> Semi-structured data has some organizational structure (tags, key-value pairs) but not a rigid schema.

3 **What is the relationship between data, information, and knowledge?**
1 Point - Only one correct choice

- ☐ **A.** Data is the application of knowledge, and information is the raw form of data.
- ☐ **B.** Information is the raw form of data, and knowledge is the processed form of information.
- ☒ **C.** Data is raw, information is processed data with meaning, and knowledge is the application of information.
- ☐ **D.** Knowledge is raw data, and information is the processed form of knowledge.

> Data is raw, information is data with context and meaning, and knowledge is the ability to use information for action.

4 **Which of the following is an example of unstructured data?**
1 Point - Only one correct choice

- ☐ **A.** A table of student grades in a CSV file
- ☐ **B.** A JSON file containing user profiles
- ☒ **C.** A collection of handwritten doctor's notes
- ☐ **D.** A relational database of customer orders

> Unstructured data lacks a predefined format or organization, such as free-text notes.

5 **Which of the following is an example of an exploratory question in data science?**
1 Point - Only one correct choice

- ☐ **A.** What was the average temperature in Paris last month?
- ☐ **B.** How many customers visited the store yesterday?
- ☐ **C.** Will the stock price of Company X increase next week?
- ☒ **D.** Is there a relationship between physical activity levels and sleep quality?

> Exploratory questions seek to uncover patterns, correlations, or relationships within the data, rather than predicting outcomes or describing past events.

6 **A dataset of public tweets about a brand is collected for sentiment analysis. Before using this data, what is a critical ethical consideration?**

1 Point - Only one correct choice

- ☐ **A.** The color scheme of the sentiment visualization
- ☐ **B.** The number of tweets collected
- ☒ **C.** Whether users provided informed consent, even if the data is anonymized
- ☐ **D.** The programming language used for analysis

> Ethical use of public social media data requires considering user consent and privacy, regardless of anonymization.

7 **Which of the following best describes tacit knowledge?**

1 Point - Only one correct choice

- ☐ **A.** Knowledge stored in databases or books
- ☒ **B.** Personal knowledge gained through experience, such as skills or intuition
- ☐ **C.** Raw data collected from sensors
- ☐ **D.** Information presented in a chart or graph

> Tacit knowledge is personal, context-specific, and difficult to formalize or share.

8 **What is a common challenge when working with unstructured data?**

1 Point - Only one correct choice

- ☐ **A.** It can be directly loaded into a relational database without preprocessing.
- ☒ **B.** It requires specialized tools (e.g., NLP, computer vision) for analysis.
- ☐ **C.** It is always stored in CSV format.
- ☐ **D.** It has a fixed schema like structured data.

> Unstructured data, such as text or images, often requires advanced techniques for processing and analysis.

9 **Why is it important to consider the source of data in data science?**

1 Point - Only one correct choice

- ☐ **A.** To choose the prettiest visualization
- ☐ **B.** To ignore external data
- ☒ **C.** To assess reliability, potential biases, and the right to use the data
- ☐ **D.** To ensure the data is always machine-generated

> The source of data affects its quality, reliability, ethical use, and potential biases.

10 **In a hospital scenario, which of the following is a privacy concern?**

1 Point - Only one correct choice

- ☐ **A.** Aggregated statistics of patient vitals
- ☒ **B.** Personal identifiers linked to health records
- ☐ **C.** Time-stamped anonymous sensor logs
- ☐ **D.** Weather data for hospital vicinity

> Linking personal identifiers to medical data raises privacy risks.

11 **Which type of analysis would help a smart city predict traffic congestion?**

1 Point - Only one correct choice

- ☐ **A.** Descriptive statistics only
- ☒ **B.** Predictive modeling using sensor and GPS data
- ☐ **C.** Data cleaning
- ☐ **D.** Creating user dashboards

> Predictive models use input signals to anticipate future congestion.

12 **Which of the following is an example of machine-generated data?**
1 Point - Only one correct choice

- ☐ **A.** Customer reviews on an e-commerce website
- ☒ **B.** Temperature readings from a weather sensor
- ☐ **C.** Handwritten notes from a doctor's appointment
- ☐ **D.** Social media posts about a new product

> Machine-generated data is produced automatically by devices or software, such as sensors or logs.

13 **A company collects employee attendance records from its internal HR system. This data is best described as:**
1 Point - Only one correct choice

- ☐ **A.** External and proprietary
- ☒ **B.** Internal and proprietary
- ☐ **C.** External and open
- ☐ **D.** Internal and open

> Employee attendance records are collected within the organization (internal) and are typically restricted to the company (proprietary).

14 **Publicly available datasets from government agencies (e.g., census data) are usually classified as:**
1 Point - Only one correct choice

- ☐ **A.** Proprietary and internal
- ☐ **B.** Proprietary and external
- ☒ **C.** Open and external
- ☐ **D.** Open and internal

> Government datasets are typically freely accessible (open) and collected outside of any single organization (external).

15 **Which of the following scenarios describes human-generated data?**
1 Point - Only one correct choice

- ☐ **A.** GPS coordinates logged by a smartphone app
- ☒ **B.** A hand-filled survey about customer satisfaction
- ☐ **C.** Automated factory sensor logs
- ☐ **D.** Web server access logs

> Human-generated data is created directly by people, such as surveys, reviews, or manual entries.