DATA SCIENCE (CDA)
CLASS ASSESSMENT 1 (UNITS 1 AND 2, MODEL A)

1. From the five "the value of data" types seen in class, which one corresponds to the following example?

An app is created to collect and share information from drivers,
and give real-time information, knowledge and advice about routes

a) That data is valuable for me (out → in)
b) My data is valuable for others (in → out)
c) New data is valuable for others (∅ → out)
d) That data is valuable for others (out → out)

2. What is the correct sequence of stages of CRISP-DM?
a) Data anonymisation, data integration, validation, modelling and revision
b) Business understanding, data understanding, data preparation, modelling, evaluation and deployment.
c) Data integration, data anonymisation, validation, modelling, application and revision.
d) Data integration, validation, data anonymisation, modelling, application and revision.

3. Which of the following visual attributes (retinal variables) is NOT suitable to encode ordinal data (for instance, "low", "medium" and "high")?
a) Texture.
b) Size.
c) Colour value.
d) Orientation.

4. What is an OLAP operator?
a) An efficient JOIN operator for NOSQL databases.
b) An efficient JOIN operator for relational databases.
c) An efficient JOIN operator for hadoop.
d) A usually interactive procedure that allows an OLAP query to be refined in terms of aggregation or the dimensions used in the query over a multidimensional schema.

5. Which of the following actions is a suitable option to handle missing values of categorical variables?
a) replace them by -1 or another impossible value.
b) replace the value by the mode.
c) exchange rows and columns.
d) replace them by 0.

6. In a dataset about tree species over the Amazon, where we have three attributes (latitude, longitude and type-of-tree), and data is sparse and unevenly distributed geographically, we want to select a representative sample that covers all geographical areas of the jungle. Which sampling method should we use to obtain such a sample?

a) Simple random sampling.
b) Stratified random sampling.
c) Group sampling.
d) Exhaustive sampling.

7. Which of the following data transformations should we apply to convert the "ethnicity" field with values {Asian, Hispanic, Black, White) into four binary fields?

a) **Numerisation**
b) Discretisation
c) Normalisation
d) Principal Component Analysis

8. Which of the following statements is CORRECT?

a) An outlier is the name we use to refer to each of the rows of a dataset.
b) **The data are unbiased if they are representative of the population of interest.**
c) The selection of some characteristics (attributes) in a dataset is called sampling.
d) The representation of information using spatial or graphical resources is called information transformation.

9. Imagine you anonymise patient data by converting ID code into new meaningless codes, and then publish the data. Which of the following statements is CORRECT?

a) People can never identify any patient.
b) People could identify the patients only if the new codes have backward traceability.
c) People could identify the patients only if the birth date is available.
d**) People could identify some patients using some other attributes.**

10. Which of the following statements is FALSE?

a) Models in data science tend to find discriminative patterns.
b) A good knowledge of the domain is the most important issue to create good derived attributes
c) Derived attributes are the attributes that do not exist in the physical database, but their values are derived from other attributes present in the database.
d) **Discretising categorical attributes is one of the techniques used to deal with outliers or missing values.**

ASSESSMENT
Answer Sheet
MODEL A

Surname:	**Name:**
Group in English: ☒	

In the following table, circle the correct answer for each question.

Question	**Answer**			
1	a	b	c	d
2	a	b	c	d
3	a	b	c	d
4	a	b	c	d
5	a	b	c	d
6	a	b	c	d
7	a	b	c	d
8	a	b	c	d
9	a	b	c	d
10	a	b	c	d

The result will be calculated by the statistical correction formula:

$$(\text{Right} - \text{Wrong}/3) \times 1$$

which discounts the probability of getting a right answer by chance on a question with four possibilities.

The mark is between 0 and 10.

Remember that this assessment is just 10% of the final qualification for the course.