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?

By collecting and sharing information from students at a university, we can predict how many will drop out next year.

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

2. What is CRISP-DM?

a) A methodology for the process of data mining and knowledge discovery from data.
b) An extraction-transformation-load (ETL) tool for data silos.
c) A database specialised for Big Data .
d) A data analytics tool.

3. Which of the following visual attributes (retinal variables) is suitable to encode ordinal data (for instance, “low”, “medium” and “high”)?

a) Texture.
b) Size.
c) Shape.
d) Colour (Hue).

4. Which of the following claims is TRUE ?

a) Outlier is the name we use for referring to each row in a dataset.
b) Data is biased if it is representative of the population of interest.
c) The selection of some rows (examples) in a dataset is called sampling.
d) The depiction of information using spatial or graphical representations is called Information Transformation.

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 an average value.
c) exchange rows and columns.
d) replace them by 0.

The question was wrong as it should be about numeric (quantitative) variables, and the answer was meant to be “b”. But for categorical variables the mean doesn’t make sense, so all answers are wrong. As a result, the question is cancelled, but I’ll give the point..

6. What is a multidimensional datawarehouse?

a) A plot with more than three dimensions that is represented using parallel coordinates.
b) A NoSQL database, where queries are performed using a multidimensional XML query language.
c) An unstructured database where data is located in different data silos.
d) A structured database where information is represented by central facts and attributes are arranged into different dimensions.

7. What is data curation?

a) The recovery of data lost in a datawarehouse or NOSQL database.
b) A term to group those technologies in big data dealing with data integration, preparation, quality, privacy, security, metadata and general manipulation of the data.
c) The transformation of the data such that individual identities are not traceable.
d) The set of technologies that support the storage of data, the data infrastructure.

8. What is the transformation of information such that the identity of people cannot be recovered?

a) Anonymisation.
b) Privacy.
c) Discrimination.
d) Personalisation.

9. If we have an attribute with five possible values of increasing value: “very low”, “low”, “average”, “high” and “very high”?

a) A numerisation “1 to n” would capture the natural gradation seen in the attribute values.
b) A numerisation “1 to 1” would capture the natural gradation seen in the attribute values.
c) A discretisation “1 to n” would capture the natural gradation seen in the attribute values.
d) A discretisation “1 to 1” would capture the natural gradation seen in the attribute values.

10. 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.

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	Cancelled
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:

(Right – Wrong/3) × 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.