**Question 1 - multiple choice, no shuffle**

Using a collection of previously labelled objects and their descriptions to attaching a label to a new object is called:

\*A: A​ classification problem

B: A​ sorting problem

C: A​ clustering problem

D: A​ stable matching problem

**Question 2 - multiple choice, no shuffle**

I​n a supervised learning problem, the labels for the "training" objects used to fit or train a model are:

\*A: K​nown

B: U​nknown

**Question 3 - multiple choice, no shuffle**

W​hich of the following stages comes latest in the data science pipeline, as introduced in this course?

A: Model deployment

B: P​roblem formulation

C: M​odel training and selection

\*D: M​onitoring

**Question 4 - checkbox, no shuffle, no partial credit**

A​t which stages in the data science pipeline can ethical risks occur?

\*A: D​ata collection

\*B: Data preprocessing

\*C: P​roblem formulation

\*D: F​eature engineering

\*E: M​odel training and selection

\*F: M​odel deployment

\*G: M​onitoring

**Question 5 - checkbox, no shuffle, no partial credit**

An adult patient dataset comprises measurements of age, gender, height, weight, as well as whether patients recovered from a serious medical condition that occurs more often in men, after being given a certain treatment. The data were collected via an anonymous survey and the patients could choose not to answer. You inspect the data and find out that the weight information was not disclosed for 30% of the patients. Which of the following mitigations might introduce ethical risk? (Note: although this question can be answered with the help of formal statistical concepts, for now just try to use common sense).

\*A: Ignore all the rows (patients) where the weight information is missing.

Feedback: If the patients who decided not to disclose their weight are not a random representative sample of the patient population, this might bias the answers in ways that disadvantage those patients. For example, if patients over or under a certain weight prefer not to disclose their weight, and the treatment fails to work for just these patients, a misleading answer could be given by ignoring these rows.

\*B: Fill in the missing values with the average value across the whole patient dataset.

Feedback: Women tend to weigh less than men, which would result in their imputed values being biased upwards. Given women are under-represented in this dataset because the disease occurs more frequently in men, this would result in considerable bias in the weight values for women vs men.

\*C: Delete the weight variable altogether

Feedback: Although simple solutions like the ones listed above incur possibility of bias and hence harm, we would have to ascertain whether more advanced methods can mitigate the risk further, and following that step, whether the residual risk is worth taking in return for the patient benefit that might result from using the weight information on the remaining 70% of patients. This will also depend on the use of the model: what decisions will be made on its basis, and what is the cost of a mistake?