Test yourself 04

Rate your confidence of being able to answer the below questions saying A, B or C, where:

- A. I am confident that I know the answer to this question
- B. I know at least 50% of the answer to this question, within 20 minutes I could find the required resources to enable a complete answer
- C. I am not confident that I can answer the question at this time.

Part I

1. PCA
2. PCA
3. PCA
4. clustering
5. clustering
6. overfitting
7. KNN
8. decision tree

Part II

1. PCA
2. PCA
3. PCA
4. clustering
5. clustering
6. Could you explain why we are using data splitting into train, validation and test in machine learning?
7. What is true about data splitting into train, validation and test in machine learning?

a) we train ML methods such as classification on train data and check models on validation and test data to assess the prediction power on the unknown data sets
b) we use validation data to check if our implementation of ML is working correctly
c) we split data to have multiple dataset to assess ML performance on
d) we split data to deal with overconfident estimation of future performance

6. Could you explain how knn classification work using example data and Euclidean distance?
7. Given data below, how would a new observation be classified assuming k = 1

```
x <- c(1, 1, 2, 4, 4, 2)
y <- c(1, 2, 2, 1, 2, 1)

data <- data.frame(x = x,  y=y)
(data)
```

x y 1 1 1 2 1 2 3 2 2 4 4 1 5 4 2 6 2 1

x	y	label
1	1	A
1	2	A
2	2	A
4	1	B
4	2	B
2	1	?