ML quiz2

**Question 1:**

**Sara wants to build cat/dog images dataset. She invited 10 friends with**

**Overall, 7 cats and 5 dogs. She took 1000 images of each pet. What do you**

**think of this dataset for machine learning?**

**Answer:**

1. **The data lake for diversity cues the data is 13k images but for only 13 animals**
2. **Since each pet appears 1000 times at the data set if we use the data to train ml model the data well make the model memorize the pet not learn because the same animal appears 1000 times but its just 13 animals we talking about so if we test the model in real life scenario will defiantly fail**

**Question 2:**

**In a volleyball classification task, 200 Olympiad videos were downloaded. For**

**every video, 100 1-min short video clips were extracted for winning and losing**

**points. Then the 20000 were shuffled and split to train/test/val. What do you**

**think of this dataset for machine learning?**

**Answer 2:**

**science we split the data after we make form each video from the 200 ones then we have leakage problem because the data from the train set will appear again in the validate and test sets and that would make as unsure about if the model memorized or learned this data**

**Questions B: True or False**

**1. If the model performs well on the test set, it suggests that it has captured the**

**underlying patterns in the training data and can make accurate predictions on**

**similar instances (False)**

**2. There is always a chance of encountering data in real-world that is**

**significantly different from the training and test sets (True)**

**3. If the training data contains biases that are not representative of the target**

**population, the model may not be able to generalize well (True)**

**4. The distribution of the data may change over time, rendering the model's**

**learned patterns outdated and affecting its ability to generalize to new data (True)**

**Questions C: True or False**

1. **The purpose of train-test splitting in machine learning is to assess the model's ability to generalize to unseen data (True)**
2. **The training set is used to optimize the model's parameters, while the test set is used to tune hyperparameters (False)**
3. **The purpose of using a separate test set is to provide an unbiased evaluation of the final chosen model's performance (True)**
4. **Randomly shuffling the data before splitting it into train, validation, and test sets helps the independency of samples (as in IID) (True)**
5. **If a model performs well on both the validation and test sets, it is guaranteed to have good performance on unseen data(False)**

**Questions D: True or False**

1. **The quality and representativeness of the training data are crucial for generalization (True)**

**2. Sometimes, a smaller but high-quality training set may outperform a larger but lower-quality training set. (True)**

**3. A larger training set will always result in better generalization performance for a machine learning model (False)**

**Question E**

**● In a computer vision pipeline, we have 3 models A, B, C**

**○ Input to A is an image of a person**

**○ Input to B is an image of a person + output ofA**

**○ Input to C is an image of a person + output ofB**

**○ Final system output is output of C**

**● Data Collection team collected 300 persons data.**

**● The data is provided for model’s owner to use for his train/val/test purposes**

**● The performance of each model can be reported separately**

**● A more important performance is for the whole system (the pipeline)**

**○ Find a possible mistake in computing the pipeline whole performance**

**Answer:**

**Because the model b take the model a output as input and the same with model c and b and the data will be passed for the model a b c then even we split the data to train val test still will have leakage problem and that’s because the model a that trained on the test and val data will passed to the model b as input and the same for c**