DAILY MACHINE LEARNING INTERVIEW QUESTIONS

1. What Are the Different Types of Machine Learning?

There are **three types** of machine learning:

- **Supervised Learning**

In supervised machine learning, a model makes predictions or decisions based on past or labeled data. Labeled data refers to sets of data that are given tags or labels, and thus made more meaningful.

- **Unsupervised Learning**

In unsupervised learning, we don't have labeled data. A model can identify patterns, anomalies, and relationships in the input data.

- **Reinforcement Learning**

Using reinforcement learning, the model can learn based on the rewards it received for its previous action.

Example

Consider an environment where an agent is working. The agent is given a target to achieve. Every time the agent takes some action toward the target, it is given positive feedback. And, if the action taken is going away from thegoal, the agent is given negative feedback.

2. What is Overfitting, and How Can You Avoid It?

The Overfitting is a situation that occurs when a model learns the training set too well, taking up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data. When a model is given the training data, it shows 100 percent accuracy—technically a slight loss. **But, when we use the test data, there may be an error and low efficiency. This condition is known as overfitting.**

There are **multiple ways of avoiding overfitting, such as:**

- **Regularization:** It involves a cost term for the features involved with the objective function
- Making a simple model. With lesser variables and
- parameters, the variance can be reduced
- **Cross-validation methods like k-folds can also be used.**
- If some model parameters are likely to cause overfitting,
- techniques for regularization like LASSO can be used that
- penalize these parameters

3. What is ‘training Set’ and ‘test Set’ in a Machine Learning Model? How Much Data Will You Allocate for Your Training, Validation, and Test Sets?

There is a three-step process followed to create a model:

- **Train the model**
- **Test the model**
- **Deploy the model**
-

Training set :

- The training set is examples given to the model to
- analyze and learn.
- 70% of the total data is typically taken as the training
- dataset.
- This is labeled data used to train the model.

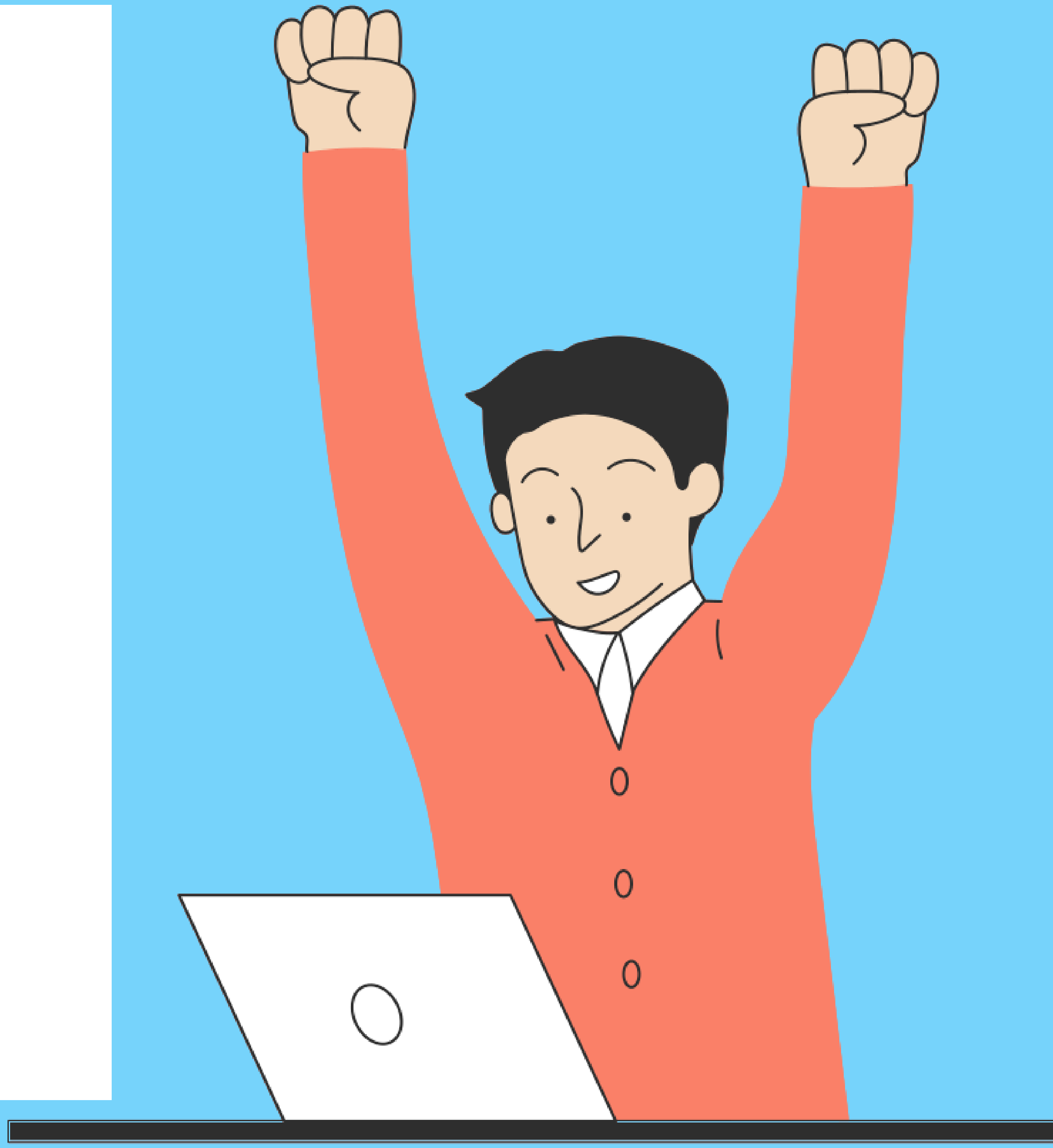
Test set:

- The test set is used to test the accuracy of the
- hypothesis generated by the model
- Remaining 30% is taken as testing dataset
- We test without labeled data and then verify results with
- labels

Consider a case where you have labeled data for 1,000 records. One way to train the model is to expose all 1,000 records during the training process. Then you take a small set of the same data to test the model, which would give good results in this case.

But, this is not an accurate way of testing. So, we set aside a portion of that data called the ‘test set’ before starting the training process. The remaining data is called the ‘training set’ that we use for training the model. The training set passes through the model multiple times until the accuracy is high, and errors are minimized.

Now, we pass the test data to check if the model can accurately predict the values and determine if training is effective. If you get errors, you either need to change your model or retrain it with more data.



Regarding the question of how to split the data into a training set and test set, there is no fixed rule, and the ratio can vary based on individual preferences.

4. How Do You Handle Missing or Corrupted Data in a Dataset?

One of the easiest ways to handle missing or corrupted data is to drop those rows or columns or replace them entirely with some other value.
There are two useful methods in Pandas:

- **IsNull() and dropna(**) will help to find the columns/row with missing data and drop them.
- **Fillna()** will replace the wrong values with a placeholdervalue

5. How Can You Choose a Classifier Based on a Training Set Data Size?

When the training set is small, **a model that has a right bias and low variance seems to work better because they are less likely to overfit.**

For example, Naive Bayes works best when the training set is large. Models with low bias and high variance tend to perform better as they work fine with complex relationships.

THANK YOU!

Have a
great day
ahead.