# Machine Learning Interview Questions Part 5

1. Is it better to have too many false positives or too many false negatives? Explain.

It depends on the question as well as on the domain for which we are trying to solve the problem. If you're using Machine Learning in the domain of medical testing, then a false negative is very risky, since the report will not show any health problem when a person is actually unwell. Similarly, if Machine Learning is used in spam detection, then a false positive is very risky because the algorithm may classify an important email as spam.

2. What Are the Applications of Supervised Machine Learning in Modern Businesses?

Applications of supervised machine learning include:

- Email Spam Detection

Here we train the model using historical data that consists of emails categorized as spam or not spam. This labeled information is fed as input to the model.

- Healthcare Diagnosis

By providing images regarding a disease, a model can be trained to detect if a person is suffering from the disease or not.

- Sentiment Analysis

This refers to the process of using algorithms to mine documents and determine whether they're positive, neutral, or negative in sentiment.

- Fraud Detection

Training the model to identify suspicious patterns, we can detect instances of possible fraud.

3. What Are Unsupervised Machine Learning Techniques?

There are two techniques used in unsupervised learning: clustering and association.

Clustering
Clustering problems involve data to be divided into subsets. These subsets, also called clusters, contain data that are similar to each other. Different clusters reveal different details about the objects, unlike classification or regression.

Association
In an association problem, we identify patterns of associations between different variables or items.
For example, an ecommerce website can suggest other items for we to buy, based on the prior purchases that we have made, spending habits, items in your wishlist, other customers' purchase habits, and so on.

4. How Is Amazon Able to Recommend Other Things to Buy? How Does the Recommendation Engine Work?

Once a user buys something from Amazon, Amazon stores that purchase data for future reference and finds products that are most likely also to be bought, it is possible because of the Association algorithm, which can identify patterns in a given dataset.

5. What Is 'training Set' and 'test Set' in a Machine Learning Model? How Much Data Will We Allocate for Your Training, Validation, and Test Sets?

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

1. Train the model
2. Test the model
3. Deploy the model

Training Set
Test 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
- 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 we 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 we 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 we get errors, we 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.