**Question 1.**

What are the three stages to build the hypothesis or model in machine learning?

**Ans.**

1. Model Building
2. Applying the model
3. Testing the model

**Question 2.**

What is the standard approach to supervised learning?

**Ans.**

We use supervised learning if we know the labels beforehand. We divide the dataset into two parts

1. Training data
2. Testing data

We do the machine learning algorithm on Training data and build the model. Based on the output of this, we test this function on Testing data. We compare between Actual and Predicted versions and check accuracy.

**Question 3.**

What is training set and test set?

**Ans.**

* Training Set –

Training Set is the set on which the machine learning model is built. We usually split using train\_test\_split in Python sklearn to split training and testing data. We need to remove any duplicates and null values before training set is fed to the algorithm

* Test Set –

The test set is a dataset used to measure how well the model performs at making predictions on that test set. We check the accuracy using score() or ROC curve, classification report and confusion matrix

**Question 4.**

What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?

**Ans.**

The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model.

Bootstrap aggregating, also called **bagging**, is a [machine learning ensemble](https://en.wikipedia.org/wiki/Ensemble_learning) method designed to improve the stability of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms used in [statistical classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis). It also reduces [variance](https://en.wikipedia.org/wiki/Variance) and helps to avoid [overfitting](https://en.wikipedia.org/wiki/Overfitting). Although it is usually applied to [decision tree](https://en.wikipedia.org/wiki/Decision_tree_learning) methods, it can be used with any type of method.

**Boosting**is a machine learning ensemble algorithm for primarily reducing bias, and variance in supervised learning, and a family of machine learning algorithms that convert weak learners to strong ones through various epochs.

**Question 5.**

How can you avoid overfitting ?

**Ans.**

There are number of ways to avoid overfitting. We can use cross-validation so that the data is well organized and multiple times fed to algorithm. We can take the average of these accuracies to find optimal performance. We can use random forests and boosting and bagging.

We can finetune the dataset for data cleaning by combining various data cleaning, transformation and creating dummy variables for categorical features etc so that the model can benefit the maximum from this.

Pruning the trees also can be beneficial when working with tree based algorithms.