**Q 1: What are the three stages to build the hypotheses or model in machine learning?**

**Ans:**

a) Model building

b) Model testing

c) Applying the model

**Q 2: What is the standard approach to supervised learning?**

**Ans:**

Labeled data is the standard approach for the supervised learning. Where, the result says that the points labeled as “F” (Failed) or “R” (Runs).

**Q3: What is Training set and Test set?**

**Ans:**

A **Training Set** is a dataset used to train a model. In training the model, specific features are picked out from the training set. These features are then incorporated into the model.

The **Test Set** is a dataset used to measure how well the model performs at making predictions on that test set.

If the prediction scores for the test set are unreasonable, we’ll have to make some adjustments to our model

and try again.

**Q4: 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. Bagging is a method in ensemble for improving unstable estimation or classification schemes.

**Bagging** builds multiple base models with re-sampled training data with replacement. We train base classifiers on different samples of training data. Using random subsets of the data to train base models promotes more differences between the base models.

**Boosting** is another ensemble technique to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors. In other words, we fit consecutive trees (random sample) and at every step, the goal is to solve for net error from the prior tree.

**Q5: How can you avoid overfitting ?**

**Ans:**

Identify the important variables and think about the model that you are likely to specify, then plan ahead to collect a sample large enough handle all predictors, interactions, and polynomial terms your response variable might require.