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

a)      Model building

b)      Model testing

c)       Applying the model

**2.) What is the standard approach to supervised learning?**

The standard approach to supervised learning is to split the set of examples into training set and test set.

**3.) What is Training set and Test set?**

In various areas of Machine Learning, a certain set of data is split in the proportion of 70% and 30% to test the accuracy of our predictions. Where our model of predicative analysis is applied on 70% of the data and we come out with the hypothesis and some function as y=f(x).

Now this y = f(x) is tested on Test data which is 30% of the entire data to test the accuracy of hypothesis generated with the training set of data.

**4.) What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?**

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to **decrease** **variance**(bagging), **bias** (boosting), or **improve predictions** (stacking).  Bagging is a method in ensemble for improving unstable estimation or classification schemes.  While boosting method are used sequentially to reduce the bias of the combined model.  Boosting and Bagging both can reduce errors by reducing the variance term.

**5.) How can you avoid overfitting?**

Following are the commonly used methodologies to avoid overfitting:

1. **Cross-Validation** : Cross Validation in its simplest form is a one round validation, where we leave one sample as in-time validation and rest for training the model. But for keeping lower variance a higher fold cross validation is preferred.
2. **Train with more data:** Training more data will helpalgorithm detect the signal more better. But if we add more noisy data then this would not help.
3. **Remove features:** Some algorithms have built-in feature selection. For those that don’t, you can manually improve their generalizability by removing irrelevant input features.
4. **Early Stopping**: Early stopping rules provide guidance as to how many iterations can be run before the learner begins to over-fit.
5. **Pruning**: Pruning is used extensively while building CART models. It simply removes the nodes which add little predictive power for the problem in hand.
6. **Regularization**: It introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce cost term.