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

The three stages to build the hypotheses or model in machine learning are .a) Model building b) Model testing c) Applying the model

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

Supervised learning is a type of machine learning that enables the model to predit future outcomes after they are trained based on the past data. To train the model, first, a set of inputs and outputs are fed to it. Say for example you want your model to be able to recognize a car from given data. Training is done by providing set of inputs and outputs that help the system understand what the essential features are, that define a car. Now, when the new data is fed to the model, the model is in a position to recognize a car. Every output that the model provides,a long with the new data that facilitated the output, becomes the new input-output combination that is fed as training data into the model for learning.

**3) What is training set and test set?**

In Machine Learning, 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. Thereby, if the training set is labeled correctly, the model should be able to learn something from these features.

The test set is a dataset used to measure how well the model performs at making predictions on that test set. If the predictions scores for the testset are unresonable, we'll have to make some adjustments to our model and try again.

In an intutive way Training set is an examples given to the learner, while Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of example held back from the learner.

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

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 both can reduce errors by reducing the variance term.

Boosting method are used sequentially to reduce the bias of the combined model. Boosting can reduce errors by reducing the variance term.

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

Here are the few solutions for overfitting:

**Pruning**

**Pre-Pruning**

In pre-pruning, it stops the tree construction bit early. It is preferred not to split a node if its goodness measure is below a threshold value. But it’s difficult to choose an appropriate stopping point.

**Post-Pruning**

In post-pruning first, it goes deeper and deeper in the tree to build a complete tree. If the tree shows the overfitting problem then pruning is done as a post-pruning step. We use a cross-validation data to check the effect of our pruning. Using cross-validation data, it tests whether expanding a node will make an improvement or not. If it shows an improvement, then we can continue by expanding that node. But if it shows a reduction in accuracy then it should not be expanded i.e, the node should be converted to a leaf node.

**Cross-validation-** We partition the data into k subsets, called folds. Then, we iteratively train the algorithm on k-1 folds while using the remaining fold as the test set (called the “holdout fold”).

**Train with more data-** It won’t work everytime, but training with more data can help algorithms detect the signal better. In the earlier example of modeling height vs. age in children, it’s clear how sampling more schools will help your model.

**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.

**Early stopping-**When you’re training a learning algorithm iteratively, you can measure how well each iteration of the model performs. Up until a certain number of iterations, new iterations improve the model. After that point, however, the model’s ability to generalize can weaken as it begins to overfit the training data. Early stopping refers stopping the training process before the learner passes that point.

**Regulization-**Regularization refers to a broad range of techniques for artificially forcing your model to be simpler. The method will depend on the type of learner you’re using. For example, you could prune a decision tree, use dropout on a neural network, or add a penalty parameter to the cost function in regression. Oftentimes, the regularization method is a hyperparameter as well, which means it can be tuned through cross-validation**.**

**Ensembling-** Ensembles are machine learning methods for combining predictions from multiple separate models. There are a few different methods for ensembling, but the two most common are a)Bagging b)Boosting.