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

* **Model building**

The first stage in any Machine Learning problem is the representation of the problem in a formal language. This is where we usually define the Machine Learning task to be performed based on the data and the business objective or problem to be solved. Usually this stage of the problem is masked as another stage, which is the selection of the ML algorithm or algorithms (you might have multiple possible model representations at this phase). When we select a target algorithm, we are implicitly deciding on the representation that we want use for our problem. This stage is akin to deciding on the set of hypothesis models , any of which can be the solution of our problem .For example, when we decide the Machine Learning task to be performed is regression looking at our dataset and then select linear regression as our regression model. Then we have decided on the linear combination based relationship between the dependent and the independent variables. Another implicit selection made in this stage is deciding on the parameters/weights/coefficients of the model that we need to learn

* **Model testing**

Once we decide on the representation of our problem and possible set of models, we need some judging criterion or criteria that will help us choose one model over the others, or the best model from a set of candidate models. The idea is to define a metric for evaluation or a scoring function\loss function that will help enable this. This evaluation metric is generally provided in terms of an objective or an evaluation function (can also be called a loss function). What these objective functions normally do is provide a numerical performance value which will help us to decide on the effectiveness of any candidate model. The objective function depends on the type of problem we are solving, the representation we selected, and other things. A simple example would be the lower the loss or error rate, the better the model is performing

* **Applying the model**

The final stage in the learning process is optimization. Optimization in this case can be simply described as searching through all the hypothesis model space representations, to find the one that will give us the most optimal value of our evaluation function

1. What is the standard approach to supervised learning?

The standard approach to supervised learning is to split the dataset into the training set and testing set.

Supervised learning is a 2 step process:

* Learning/Training – Learn a model using the training data
* Testing – Test the model using unseen test data to assess the model accuracy

1. What is Training set and Test set?

* Training 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

* Test set

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.

1. 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 - (Bootstrap Aggregating)  is an parallel ensemble method. First, we create random samples of the training data set (sub sets of training data set). Then, we build a classifier for each sample. Finally, results of these multiple classifiers are combined using average or majority voting. Bagging helps to reduce the variance error.Eg- Random Forest
* Boosting - sequential ensemble: try to add new models that do well where previous models lack. aim to **decrease bias**, not variance. suitable for low variance high bias models. example of a tree based method is **gradient boosting,Adaboosting**

1. How can you avoid overfitting ?

There are different solutions to avoid overfitting :

* Cross-validation - Use your initial training data to generate multiple mini train-test splits. Use these splits to tune your model. Eg – K-fold cross validation
* Train with more data
* Remove irrelevant input features
* Early stopping - stopping the training process before the learner passes that point
* Regularization - refers to a broad range of techniques for artificially forcing your model to be simpler.
* Ensembling - Ensembles are machine learning methods for combining predictions from multiple separate models