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

Solution:

- **Model building:** We apply a fraction of the pre-processed data on a suitable machine learning algorithm. The algorithm will train itself by that data and then builds a model
- **Model testing:** After model formation we apply unseen data to the model to test the model.
- **Applying the model:** After building and testing the model we apply that model in the real world problems.

2. What is the standard approach to supervised learning?

Solution:

In supervised learning we get labeled data for building the model. So first we split the data in training and testing dataset and then build model on training data and apply testing data on model to test. We'll test the model by comparing the predicted value by the model to the labeled values of data.

3. What is Training set and Test set?

Solution:

When we get a data set we divide them into 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. Thereby, if the training set is labeled correctly, the model should be able to learn something from these features.

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.

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

Solution:

Ensemble learning method: It means grouping multiple week learning models to form a strong model so that we can obtain better prediction.

Bagging: It is uses in Low Bias - High Variance problem. This problem occur when the model over fits like in Decision tree. In this case various models are built in parallel and each model gets trained on randomly selected samples. Then the various models vote to give the final prediction. Predictions will be averaged to get the final prediction (in case of regression) and in case of classification the final prediction will be the mode of the predicted ans.

Boosting: It is uses in Low variance - High Bias problem. This problem occurs when model under fits. In this method we first sampled the input data to generate a set of training data. Then we run an algorithm on this training data to get a trained model. Then we take all our training data to test the model and we are going to discover that some of the points are not well predicted. Now we have to build the second bag of sampled data. In this also the data will be randomly chosen, but now each data point is weighted according to error found in last model. So these values are more likely to get picked in this bag than any other data. Now we'll build a model for this sample set also then we'll test it. Here the testing will be performed on both of the model and the result will be mode of the both result (in case of classification) in case of regression result will be mean of the both result. Now again we'll find some values which are not predicted well. So we'll build one more bag and one more model and this process will continue.

5. How can you avoid over fitting?

Solution:

When we train the model on training set to such a level that it starts predicting noise or outlier present in training data correctly, then we say that model gets over fit. In this case model will show 100% accuracy for training set but its accuracy will be low on test data. There are multiple methods by which we can avoid over fitting:

Ensemble method: It is the best method to avoid over fitting and increasing accuracy. Ensembles are machine learning methods for combining predictions from multiple separate models. There are a few different methods for ensemble, but the two most common are:

1. **Bagging** attempts to reduce the chance over fitting complex models.
    - It trains a large number of “strong” learners in parallel.
    - A strong learner is a model that's relatively unconstrained.
    - Bagging then combines all the strong learners together in order to “smooth out” their predictions.
2. **Boosting** attempts to improve the predictive flexibility of simple models.
    - It trains a large number of “weak” learners in sequence.
    - A weak learner is a constrained model (i.e. you could limit the max depth of each decision tree).
    - Each one in the sequence focuses on learning from the mistakes of the one before it.
    - Boosting then combines all the weak learners into a single strong learner.

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.

Cross validation: Use the training data to generate multiple mini train-test splits and then use these splits to tune your model. Cross-validation allows you to tune hyper parameters with only your original training set. This allows you to keep your test set as a truly unseen dataset for selecting your final model.